# A MACRO-STOCHASTIC APPROACH TO IMPROVED COST ESTIMATION FOR DEFENSE ACQUISITION PROGRAMS

**THESIS**

**Allen J. DeNeve, Captain, USAF**

**AFIT-ENV-14-M-20**

**DEPARTMENT OF THE AIR FORCE**
**AIR UNIVERSITY**
**AIR FORCE INSTITUTE OF TECHNOLOGY**

**Wright-Patterson Air Force Base, Ohio**

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.  This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENV-14-M-20

# A MACRO-STOCHASTIC APPROACH TO IMPROVED COST ESTIMATION FOR DEFENSE ACQUISITION PROGRAMS

THESIS

Presented to the Faculty

Department of Systems and Engineering Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Systems Engineering

Allen J. DeNeve, BS

Captain, USAF

March 2014

AFIT-ENV-14-M-20

# A MACRO-STOCHASTIC APPROACH TO IMPROVED COST ESTIMATION FOR DEFENSE ACQUISITION PROGRAMS

Allen J. DeNeve, BS

Captain, USAF

Approved:

//signed// 6 March 2014

Erin Ryan, Lt Col, USAF, PhD (Chairman) Date

//signed// 6 March 2014

Jonathan Ritschel, Lt Col, USAF, PhD (Member) Date

//signed// 6 March 2014

Christine Schubert Kabban, PhD (Member) Date

AFIT-ENV-14-M-20

## Abstract

Inaccurate cost estimates are a recurrent problem for Department of Defense (DoD) acquisition programs, with cost overruns exceeding billions of dollars each year. These estimate errors hinder the ability of the DoD to assess the affordability of future programs and properly allocate resources to existing programs. In this research, the author employs a novel approach called "macro-stochastic" cost estimation for significantly reducing cost estimate errors in Major Defense Acquisition Programs (MDAPs). To achieve this reduction, the author first extracts and catalogs key programmatic data from 936 Selected Acquisition Reports. The author then analyzes historical trends in the data using mixed-model regression with high-level descriptive program parameters. Based on these trends, the model is found to reduce estimate errors by 18.7 percent on average, when applied to a randomly selected, historical cost estimate. However, the model is most beneficial when applied early in program life; when applied to the first cost estimate of each program in the database, the macro-stochastic technique reduces cost estimate error by over one-third. This statistically and economically significant reduction could potentially allow for reallocation of $6.25 billion, annually, if applied consistently to the DoD's portfolio of MDAPs.

# Table of Contents

Page

# List of Figures

Page

## List of Tables

## List of Equations

# A MACRO-STOCHASTIC APPROACH TO IMPROVED COST ESTIMATION FOR DEFENSE ACQUISITION PROGRAMS

## I. Introduction

### General Issue

The Department of Defense is operating in an increasingly constrained fiscal environment. In this climate of conservation and reduction, the Office of Management and Budget shows that the inflation-corrected defense budget has been reduced by approximately 17 percent since 2010 (The White House, 2014). Sequestration measures have forced the DoD to cut over $41 Billion in the last six months of 2013 (OSD Comptroller, 2013). Research, Development, Test and Evaluation outlays have fallen more than 18 percent over the last four years, proving that the acquisition budget is not shielded from these cuts (The White House, 2014). Despite these reductions, the nation still relies upon the military to produce effective weapons systems at a fair cost.

Accurately estimating the final cost of these weapon systems is difficult, largely due to the uncertainty involved. This uncertainty is an inherent part of defense acquisition due to the novelty and complexity of producing unprecedented military capabilities. Requirements instability and political considerations add to this uncertainty. It is not surprising, then, that inaccurate estimates are a constant companion to such acquisition efforts. A Government Accountability Office (GAO) study from 2012 showed that the DoD acquisition portfolio exceeded its baseline cost estimates by over $74 Billion in that year alone, an amount that would have paid for the recent sequestration cuts nearly twice over (GAO, 2012a). Such large overruns do not cultivate trust in the defense acquisition system, with Congress or the public.

Government agencies and independent organizations have conducted myriad studies to determine the major sources of cost overruns in acquisition programs and many of the suggestions resulting from these studies have been implemented (Kadish, 2005). However, these initiatives are largely aimed at reducing the aforementioned uncertainty by improving the Defense Acquisition System (DAS). The most recent of major acquisition reforms is the Weapons Systems Acquisition Reform Act of 2009, and this legislation is largely aimed at taming uncertainty in DoD acquisition. It mandates several industry best practices such as systems engineering activities and technology maturity assessments in all stages of development. The Defense Acquisition Guidebook declares that these activities are critical for managing uncertainty, and emphasizes the importance of "sufficient knowledge to reduce the risk associated with program initiation, system demonstration, and full-rate production" (DAU, 2013:906).

Managing uncertainty to reduce unforeseen program costs is one way to prevent cost growth; however, this is not the only solution. Another solution is to focus on informing better resource allocation decisions from the outset. In one report, the GAO stated that the "DoD's inability to allocate funding effectively to programs is largely driven by the acceptance of unrealistic cost estimates and a failure to balance needs based on available resources" (GAO, 2008:3). A method to improve this resource allocation is to embrace the uncertainty that typifies DoD acquisition in order to provide a more accurate initial assessment of final program cost. This research employs a technique, known as *macro-stochastic estimation*, that uses statistical methods to predict program cost estimation performance, in the earliest phases of their development, by associating them with past programs. This methodology encompasses known major cost drivers such as changes to the Acquisition Program Baseline (APB) that are categorically excluded from even the most rigorous estimates (Ryan et al., 2013).

*Cost growth* is a term frequently used to define the deviation of program cost from some baseline estimate. While this term typically connotes positive deviations (i.e., baseline estimate is lower than actual cost), negative deviations (that is, overestimates) are also included in the definition. Both types of deviations result in inefficient allocation of vital resources, and distort assessments of program affordability. Fundamentally, cost growth is based on just two elements: the initial cost estimate, and the deviation from this estimate over time. While these elements are functionally related, neither the accuracy of the initial estimate nor the total cost deviation can be known until the program is complete. The consequence of this fact is that the value of an accurate cost estimate steadily diminishes as the program matures, until the program is complete and the estimate no longer has any value.

This phenomenon of decreasing utility calls into question a popular method of coping with cost growth, which is to continually revise the estimate and generate new program baselines once overruns and other programmatic changes become apparent. The new estimates succeed in generating a more accurate picture of program cost, but since many of the programming and technical decisions will have already been made, these revised baselines possess decreasing utility.

## Problem Statement

Current cost estimates generated by independent estimation techniques are limited by their restriction to the APB. Program changes to key parameters (such as duration and procurement quantity) usually result in a revision to the baseline and the cost estimates, but historical trends in these program characteristics are not taken into account when estimating the program’s cost.  These limitations result in high acquisition cost growth relative to the original estimate that reduces the efficiency of DoD resource allocation.

## Research Objective

The objective of this research is to assist resource allocation and affordability assessments of top-level decision makers early in the life of major defense acquisition programs by providing a more accurate prediction of final acquisition costs.  This objective is accomplished by identifying general programmatic factors and trends that are correlated with acquisition cost growth in a selected subset of DoD acquisition programs, and quantifying the influence of these factors.  Factors that are initially available, such as branch of service, type of program, and amount of funding may then be incorporated into a model to predict cost growth in future programs.

## Investigative Questions

This research objective can only be accomplished once several key investigative questions are answered.

1. **What program characteristics are the most significant predictors of acquisition cost growth?**  With relatively few data points, this analytic effort seeks to achieve the best possible predictive capability using the fewest number of significant predictors.  The predictors that are the most highly correlated with acquisition cost growth patterns in programs, or in groups of programs, are incorporated into mathematical models of cost growth.

2. **How can the selected factors be used to modulate the acquisition cost estimate, and thus reduce the error?**  Two models are constructed.  The factors in the first model describe the cost growth of existing programs using all information readily available during their acquisition phase.  The second model

uses only the factors that are known at program initiation to predict the eventual, final acquisition cost of future programs. This accuracy is demonstrated through a validation of the predictive model.

3. **What level of confidence is achieved by predicting acquisition cost growth using significant factors that are available at program initiation?** Confidence interval estimation is used to assess prediction accuracy and usefulness.

## Research Focus

The intent of this research is to include as many DoD programs as possible in order to maintain relevance for the widest possible spectrum of acquisition portfolio managers. However, limitations on data collection and homogeneity, discussed in greater detail below, have confined this research to Major Defense Acquisition Programs (MDAPs) with a program initiation date of 1987 or later. Additionally, programs must have procured at least 25 percent of planned quantities, and be over 50 percent expended. These filtering criteria yield 70 programs with 937 program-years of acquisition cost data.

It is important to note the fundamental purpose of this study. Current cost estimation techniques require the use of a formal program baseline; estimators are prohibited from taking into account changes to this baseline. Therefore, the cost estimation techniques presented in this research are not intended to directly assist the acquisition program manager, or prescribe corrective action of any kind. Rather, this study is intended to provide high-level acquisition executives (such as the Milestone Decision Authority, acquisition portfolio managers, and independent cost estimating entities) with a reasonable expectation of how an entire portfolio of related acquisition programs will perform, on average, in terms of eventual cost growth.

## Methodology

The initial phase of this study involves acquiring data on defense acquisition programs through the use of Selected Acquisition Reports (SARs). Once an initial examination is conducted to identify SARs that meet the selection criteria, these data are added to an existing database on MDAP cost. Once all cost data are converted to a common Base Year within each program, the resulting database is verified for accuracy and consistency before proceeding with analysis.

Next, statistical methods are employed to determine trends in estimate errors. Since acquisition data are available for the same program across multiple years, this analysis constitutes a longitudinal study that requires modeling techniques capable of handling this type of data. Major predictors of variance are identified and used to build a model of cost growth using high-level programmatic attributes. Predictions are analyzed for robustness using confidence intervals to verify real-world utility. Finally, the model is validated using a modified cross validation technique, and the resulting predictions are used to correct the cost estimate error of each program. The reduction in estimate error is reported as the model's primary performance metric.

## Assumptions and Limitations

Due to limitations in the reporting of the SARs, and logistical considerations for this study, only MDAPs with a program initiation date later than 1987 that have completed an acceptable percentage of their acquisition are included. Only SARs are used for cost estimate data in this study, and only unclassified data are used, since consistency of reporting and ease of data aggregation are crucial to completion in the requisite research time.  Reporting standards are such that programs are not required to generate a SAR once they reach 90 percent expended or 90 percent delivered; therefore, the final estimate for a program is assumed to be the actual value for all parameters. Finally, it is assumed that the sources of acquisition cost estimate error have remained fundamentally unchanged since 1987, and that the trends and cost drivers in these programs will continue to pervade future acquisition efforts.

### Implications

While there have been many studies of acquisition cost growth in DoD programs, these have been largely diagnostic in nature—they seek to analyze and correct the source of cost overruns.  However, an accurate model of cost based upon program attributes may be prognostic.  That is, the prediction of error and uncertainty in future acquisition programs may be used to produce more realistic estimates of program cost, and may greatly aid the DoD in assessing the affordability of its most expensive acquisition efforts.

# II. Problem Background and Relevant Literature

## Chapter Overview

This chapter provides relevant background information on DoD cost estimating practices and affordability analyses in order to establish the utility of this research. This chapter then describes the contents of Selected Acquisition Reports (SARs). Results of previous SAR analysis efforts are used to inform this research, while differences are highlighted to distinguish this effort from previous SAR and acquisition cost studies. Next, this chapter presents an overview of the foundational work on macro-stochastic cost estimation techniques. Finally, commonly cited pitfalls in SAR analysis are summarized and discussed.

## Major Acquisition Program Cost Estimating Process

Section 3.4 of the Defense Acquisition Guidebook (DAG) summarizes the cost estimation and reporting process for MDAPs. The program manager for the acquisition program is responsible for preparing the *Component Cost Position* for each major milestone review. This cost position—an estimate of the program's life cycle cost—is submitted to the DoD-level cost oversight organization, the Cost Assessment and Program Evaluation (CAPE). The CAPE conducts an Independent Cost Estimate (ICE) and submits this estimate, along with their assessment of the Component Cost Position, to the Milestone Decision Authority (MDA). The MDA is responsible for assessing the quality of a program's cost estimates before certifying that program as an official acquisition Program of Record. This certification occurs at Milestone B, though a new cost estimate is accomplished at each major milestone. The MDA mediates any discrepancies between the program office and CAPE estimates, and approves a unified cost estimate for the program, called the *Service Cost Position*. This estimate forms the basis of the data provided to Congress in the SAR (DAU, 2013).

The program office estimate and ICE generation processes are rigorous, "require a large team, and may take many months to accomplish" (GAO, 2009:34). The GAO Cost Estimating Guide explains that the "key to developing a credible estimate is having an adequate understanding of the acquisition program" (GAO, 2009:57) as defined in the APB, and that this APB is generated using the "best available information at any point in time" (GAO, 2009:58). The guide emphasizes that accounting for cost risk and estimate uncertainty are crucial components of a quality cost estimate, though these components are not included in the final budget for the program.

The MDA, in addition to certifying that a program is ready for the next phase of development, must also certify that the funding requirements for this program fit within the expected future resources in the DoD's budget (GAO, 2009). This constraint is called *affordability*. The DAG clarifies the intent of the affordability assessment:

> Affordability analysis and constraints are not intended to produce rigid, long-term plans. Rather, they are tools to promote responsible and sustainable investment decisions by examining the likely long-range implications of today's requirements choices and investment decisions based on reasonable projections of future force structure equipment needs… (DAU, 2013:3.2.1).

This definition illustrates the utility of a tool, with which the MDA might determine these so-called "reasonable projections" (DAU, 2013:3.2.1) of future resource requirements—and therefore costs—of a program. Such a tool would need to be unconstrained by the

APB since the baseline is subject to change in accordance with “long-range implications of today’s requirement choices” and “future force structure equipment needs” (DAU, 2013:3.2.1).

## Contents of a SAR

Since 1969, Congress has required that MDAPs report program status on a yearly basis using the SAR (GAO, 2012b:1). These reports contain standardized data in a format specified by Title 10 of U.S. Code, section 2432. SARs may be available for a program in some cases before Milestone B, and are required until a program has expended 90 percent of its funding, or has procured 90 percent of its planned units. Unclassified SARs generated later than 1997 are available electronically in the Defense Acquisition Management Information Retrieval (DAMIR) system (Defense Acquisition Management Information Retrieval System, 2014). SARs generated prior to 1997, as well as classified SARs, have been made available to the Air Force Institute of Technology[1]. The requirement to deliver an annual SAR was only levied on MDAPs, defined as:

> Those estimated by the Under Secretary of Defense for Acquisition, Technology and Logistics to require an eventual total expenditure, including all planned increments, of more than… approximately $509 million for research, development, test, and evaluation, based on fiscal year 2010 dollars), approximately $3.054 billion for procurement, based on fiscal year 2010 dollars, or are designated as a major defense acquisition program by the Milestone Decision Authority (GAO, 2012b:2).

[1] In circumstances where program cost data is unclassified, these data may be admitted into the dataset. No classified information is present, either in this document, or in the dataset used for analysis.

These SARs are usually delivered in December of each year, though a significant threshold breach requires an interim SAR. Also, since the SAR is produced in conjunction with the President's budget, the presidential election years of 2000 and 2008 resulted in no SARs, other than those required due to a breach.

The SAR includes key programmatic information, such as staff contact information, mission descriptions, key performance parameters, procurement quantity, and schedule information. However, the bulk of the document is concerned with the cost of the program. Several key cost metrics are reported:

- Total Acquisition Cost, broken down by appropriation
- Funding profile, by appropriation
- Unit Cost, reported as Average Procurement Unit Cost, and Program Acquisition Unit Cost
- Variance from the previous SAR and from the current baseline
- Operating and Support Costs

The utility of the SARs, and the reason for their frequent use in acquisition analyses, is that they report program characteristics in a consistent manner across programs, and largely across years. There are a few notable exceptions to this consistency. SARs produced prior to 1992 typically have costs reported only in the purchasing power of the current year, whereas later SARs correct this amount to a common year. Additionally, some programs have a unique structure that requires a deviation from the standard SAR reporting format. These deviations are discussed in greater detail in the *Challenges* section, later in this chapter.

The DAG states that cost estimators "are required by Congress to report certain elements of program cost risk for MDAP and MAIS programs" (DAU, 2013:115). It further stipulates that these risk elements result in the generation of a confidence level in the cost estimate, and that this confidence level must be reported in the SAR. However, the formal legislation governing SAR reporting includes no such stipulations, and confidence levels are not reported in the SAR.

## Database Formation

SARs contain hundreds of metrics pertaining to acquisition program performance; however, these data are not in an easily-compiled format. As a result, SAR analysis requires extracting relevant data from these acquisition reports and placing them in an easily interpretable format. RAND research since 1993 has been conducted using their constantly growing SAR database, dubbed the Defense System Cost Performance Database (DSCPD). This database includes SARs from reporting programs—MDAPs, Major Automated Information Systems (MAISs), and some programs specially identified by Congress as special interest programs. A report on the DSCPD explains, "This database includes cost growth data derived from information in Selected Acquisition Reports (SARs), as well as a range of potential explanatory variables that include cost, schedule, and categorical information" (Jarvaise et al., 1996:iii). For example, the DSCPD places programs into one of the following categories: Aircraft, Helicopter, Missile, Electronic, Munitions, Vehicle, Ship, Space, and Other. Other summary-level variables include service component, contractor, prototype, precedent, and modification variables.

Since the DSCPD is designed to be the canonical database for analyses within the RAND Corporation, it includes all data points possible and continues to grow year to year. However, individual studies using these data often place completion criteria on programs allowed into the study. In a 1993 RAND study, the authors state, “Additionally, we have used only programs that have progressed three or more years past [Engineering and Manufacturing Development] start, a cutoff point that reasonably corresponds with the availability of good quality information” (Drezner et al., 1993:xii). That study also included only completed programs (90 percent expended or procured). These filtering criteria admitted only 150 of the 244 programs into the study; however, it helped ensure that the inferences and conclusions were supported by quality data. Since the current research effort involves collecting new data, such filtering criteria will be crucial to scoping the effort and ensuring quality results.

Cost data for MDAPs are also available in the form of the constituent contracts, catalogued in the Defense Cost and Resource Center (DCARC). These cost estimates are generated by contractors, not by the program office, and in some cases, they may differ from the program office estimates by substantial amounts. This discrepancy is typically worse on programs with erratic SAR estimates and large estimate errors. While the contractors’ final cost for a program should match the figure from the program office (since it is no longer an estimate), the Contractor Cost Performance Report (CPR) database does not contain the final values of all the independent variables available from the SAR. For example, the *Cost Variance Due to Economic Factors* is a metric that is reported in each SAR, but is not reported in the CPR. The true final cost of the program would still be valuable for predicting the final cost as a function of early program indicators, but a 2005 study shows that the final program cost is well approximated by estimates at 92.5 percent completion (Tracy, 2005). Since SARs report program status to 90 percent complete (and often beyond 90 percent, due to the annual report cycle), this final estimated cost is expected to adequately approximate the true final cost of the program. Estimate volatility late in program life is examined in Chapter 5 to support this assumption. Future studies may revise the model presented here by ensuring the final cost estimates are, indeed, accurate.

### Macro-Stochastic Estimation

This study is a direct follow-on to work performed at the Air Force Institute of Technology by Dr. Erin Ryan (Ryan et al., 2013). His research focuses on valuing flexibility in DoD acquisition programs using expected Life Cycle Cost (LCC) as a means of discriminating between design options with varying flexibility. Ryan's investigation of LCC estimate accuracy concludes that current acquisition reporting practices provide a poor estimate of LCC, largely due to the constraint to the static baseline:

> If, in fact, historical LCC estimates are highly inaccurate, then there may be a fundamental flaw in the traditional estimating methodology. This led to the hypothesis that long-term DoD cost estimates tend to be so poor because *they are constrained by a static APB* [emphasis in original] (Ryan, 2012:144).

Ryan proposes a methodology for decoupling estimates from the APB by predicting program errors using top-level variables that characterize the program. This methodology, which he dubbed "macro-stochastic" cost estimating, essentially "models the error in the program estimate as a random variable whose value is determined by a salient group of top-level program summary indicators" (Ryan, 2012:148). The dependent variable in Ryan's study is cost estimate error, which he derives from program estimates for the LCC.

Ryan did not use the RAND database to complete his study of LCC, as he determined that certain key aspects of the data were missing, insufficient, or difficult to use with available statistical tools. Rather, he created a new database to support his research. Since Ryan's research focuses mainly on LCC, his dataset requires that MDAPs have sufficient O&S cost estimate data. However, SARs were not required to include these data until 1985 (U.S. House of Representatives, 1984), and most did not comply until about 1990 (Hough, 1992). As a result, the only SARs with consistently reported O&S cost estimates are from 1990 and later. This span provides only 20 years for a program to complete its life cycle, thus allowing accurate estimation of the actual LCC in order to calculate cost growth. Ryan's dataset consists of 470 SARs describing 36 MDAPs, spanning 1987 to 2010. His dataset combines some categories listed as binary variables in the RAND dataset into new categories. For example, *modification* is one possible value in the *Iteration* variable; other possibilities include *new* and *variant*. However, the smaller dataset precludes the ability to use such numerous system type categories, and programs are assigned to one of four: *Aviation*, *Munition*, *Maritime*, and *Other*. Ryan's dataset is modified and expanded using the filtering criteria described above to fit the research objectives of the current effort.

By necessity, any dataset constructed to analyze SARs will include repeated data points from the same program collected across many years. This continuity violates a key statistical assumption for typical regression models, since the observations cannot be assumed to be independently distributed. Therefore, Ryan uses a mixed-modeling technique to analyze the data (Ryan et al., 2013). This technique requires the use of sophisticated modeling software, and obviates the use of many of the convenient routines that select regression model parameters automatically. Ryan validates his prognostic model using a modified Leave One Out Cross Validation (LOOCV), discussed in greater detail in Chapter 3.

Many of Ryan's techniques are adopted for this research effort, although the difference in the type of cost estimate (that is, acquisition cost versus LCC) allows significantly more observations in the dataset for this study. Furthermore, Ryan's research does not include confidence intervals on the model-corrected values, and does not perform model adequacy checking for statistical assumptions. These activities are incorporated into the current research effort.

No studies other than Ryan's have applied a macro-stochastic approach (or anything appreciably similar) to improving DoD cost estimates. However, other researchers have proposed methods for improving early program cost estimates by incorporating high-level program cost drivers. Carnegie Mellon's Software Engineering Measurement and Analysis (SEMA) Cost Estimation Research group published a study in 2011 in which they proposed a method called QUELCE, which stands for Quantifying Uncertainty in Early Lifecycle Cost Estimation (Ferguson et al., 2011). This method requires convening a panel of experts, and using their feedback to determine underlying cause-and-effect relationships that drive cost variability throughout program life. This feedback is used to construct a Bayesian Belief Network, and Monte-Carlo simulation is used to simulate possible trajectories in program estimates. In contrast, Ryan's macro-stochastic technique does not require input from experts, and allows a much more expeditious and data-driven assessment of likely baseline deviations.

### Other Notable SAR Analysis Methodologies

While macro-stochastic cost estimation may be a novel technique, SAR analysis is not. The RAND corporation has conducted many studies analyzing SAR data to "quantify the magnitude of weapon system program cost growth [and] identify factors affecting cost growth" (Drezner et al., 1993:xi). This 1993 study states that "SAR data are the basis of cost growth studies both in and out of DoD" (Drezner et al., 1993:8). Many SAR-based studies use similar methodologies in estimating cost growth. The aforementioned 1993 Drezner study, along with more recent studies in 2006 and 2007, are among the most rigorous and complete SAR analyses. Their similarities to the current research effort necessitate an examination of the methodologies they used to estimate cost growth.

Drezner, et al., state that, "Cost growth can be defined simplistically as the difference between estimated and actual costs. The direction of error measured from the estimate baseline can be either to initially understate costs, in which case cost growth occurs, or to overstate costs, in which case a cost reduction is realized" (Drezner et al., 1993:1). This difference between estimated and actual costs is frequently reported as a *Cost Growth Factor* (CGF) where values greater than one indicate actual cost greater than what was estimated (that is, an overrun), and values between zero and one indicating actual cost less than what was estimated (that is, an underrun). Analysts typically correct the CGF to account for different phenomena, the most prevalent being inflation and procurement quantity changes.

Inflation correction is a well-established technique that is often performed during SAR generation.  SARs report costs in Then Year amounts as well as corrected to a common baseline year, called Base Year costs.  The Base Year estimate uses published inflation indices to correct the dollar amount in the estimate to the purchasing power of some other year (typically, the year of the APB).  This correction allows direct comparison of cost estimates made in different years.  However, difficulty arises when the analyst wishes to directly compare costs in two different base years, either across programs, or even within a single program.  In order to preserve continuity and correctly calculate the cost estimate error, each estimate for a given program is corrected to a common Base Year, though this Base Year varies across programs.

Quantity normalization is also applied in the most rigorous SAR analyses and may be accomplished by one of several different techniques.  The RAND study from 2006 uses C*umulative Average Cost Improvement Curves* (CIC) (Arena et al., 2006) to normalize the initial estimate to the final quantity, while other studies simply track the cost variance due to quantity changes, as reported in the annual SAR.  The premise for this specific normalization (other parameters, such as engineering changes, are not normalized) is that quantity changes are outside the control of the program manager. This distinction is indicative of the underlying purpose of many SAR analyses: to search for causes of cost growth in order to inform corrective actions.  Drezner explains that:

> Nominal [unadjusted] cost growth is an appropriate measure if the only concern is the impact of cost growth on the federal budget. Adjusted [corrected for inflation and quantity] cost growth, however, is a more relevant measure when trying to determine how well program management has done in estimating and controlling costs within its command (Drezner et al., 1993:10).

Quantity changes are frequently identified as one of the most significant contributors to cost growth. The 1993 RAND report states that "Inflation and quantity are shown to have the largest effect on cost growth: the average cost growth for 125 programs after normalization is 42 percentage points lower than the unadjusted result" (Drezner et al., 1993:21). Therefore, normalizing for quantity change obfuscates one of the most powerful predictors of actual cost growth. While this effect may be out of the program manager's control, it is certainly relevant to anyone directly concerned with the federal budget.

None of these research efforts, other than Ryan's, use a mixed-modeling approach, although one 2007 study uses a dynamic panel approach "which includes cross section fixed effects…since there are clearly service specific characteristics" (Smirnoff and Hicks, 2007:9). The 2007 Smirnoff study is unique in two respects. First it uses a statistical technique that attempts to resolve subject-specific effects—in this case, service-specific—rather than simply estimating the average, and it reports the confidence in the findings. Second, it attributes cost growth to macroeconomic factors, such as war, defense budgetary trends, and acquisition reforms. These factors are not typically considered in acquisition analyses, though authors sometimes refer to specific phenomena, such as the Reagan build-up (Drezner et al., 1993:8). It is notable that the dynamic panel technique employed in the Smirnoff study requires a specified covariance structure, and the first-order autoregressive structure is selected to model the dependence within the data.  This is the same structure selected by Ryan, and is one of the candidate structures for the current research effort.

## Challenges in SAR Analysis

Several pitfalls and inherent difficulties exist when using SAR data to analyze acquisition cost.  All of the authors discussed above point out limitations to using their dataset.  Paul Hough, one author of the 1993 RAND study, wrote a separate report, the sole purpose of which is to identify such pitfalls and urge caution when interpreting SAR analysis results.  This section provides an overview of the challenges listed in that report, and in other relevant SAR analysis reports.  Similar challenges listed by different authors are grouped into categories below.  While some of these challenges are endemic, and shared by the current research effort, the list of assumptions and limitations pertaining to the current effort is discussed in Chapter 3.  The impact of the applicable challenges and assumptions are discussed in Chapter 5.

**Pitfall 1: Omission of major cost elements.**  Exclusion or obfuscation of significant cost elements can diminish the apparent size of a program (Hough, 1992). Also, program managers will frequently establish a margin for error in the budget; this practice will inflate the apparent size of the program, but can deflate apparent cost growth.  Jarvaise remarks on this fact when creating his SAR database: “Unfortunately, SARs do not reveal the amount allocated as a management reserve. Since the amount of contingency funds cannot be separated from the total funding for each program, the impact of these funds cannot be estimated” (Jarvaise et al., 1996:7).  Therefore, it is important to note that the current research effort is measuring the error in the official cost estimates provided to Congress, not necessarily the estimates endorsed internally by the program.

Contractor-borne costs, such as the expenditures during preliminary research and development efforts, and overruns in Firm-Fixed-Price (FFP) contracts, are not reported in the SAR (Hough, 1992; Jarvaise et al., 1996). Hough explains that technical deficiency is an unaccounted source of cost, since the price of bringing a deficient system up to the promised capability is not estimated when such deficiencies occur. This is a tenet of modern Earned Value Management (EVM), where the Budgeted Cost of Work Performed (BCWP) must be compared to the Actual Cost of Work Performed (ACWP) to determine loss of *value* in a program. The key distinction between O&S costs, management reserve, and all other omissions, is that these first two are costs directly incurred by the government. These EVM principles are irrelevant to this effort, which focuses only on the actual dollars required to fund the programs.

**Pitfall 2: Changes to reporting requirements and guidelines.** Major revisions to SAR reporting requirements cause discontinuity and disparities over time that make it difficult to compare early estimates to actual expenditures. Hough reports that from February of 1968, to June of 1989, DoD Instruction 7000.3 (the instruction pertinent to SAR generation) underwent sixteen revisions, an average of one per year (Hough, 1992). These changes ranged from mandating cost reporting in Base Year dollars, to major restructuring of cost-variance categories. The restructuring of O&S cost categories proves especially problematic since there is not any way to divide early program costs (reported in nine categories) into the newer seven-category system (Hough, 1992). Even the threshold for being classified as an MDAP has changed at least five times, from $25 million RDT&E and $100 million Procurement in Then Year 1969 dollars, to the current threshold of $509 million RDT&E and $3.1 billion in procurement (BY 2010 dollars) (Drezner et al., 1993; GAO, 2012b). Such changes prove problematic to any SAR analysis effort, as the definitions used to categorize the programs are inconsistent across time. Even when reporting requirements are stable for any length of time, the GAO reports that many of these requirements are not followed (GAO, 2012b).

**Pitfall 3: Confusing program structure and changes to that structure.** Hough provides three concrete examples of programs that, through restructuring, took on dramatically higher or lower costs than were initially estimated. However, he points out, these changes were due to combining of one program under another, the cancellation and subsequent re-start of a program, or other large dissimilarity with the initial baseline. Such changes are relatively common for long running programs where initial acquisition initiates a new block buy before the previous production is terminated.

When such rifts are encountered in program continuity, it is often impossible to extricate the sources of cost. Such programs must often be omitted from the database entirely. It is also difficult to account for cost growth in programs where costs are split across multiple services. Hough provides an example of how the AMRAAM program saw cost growth in the Air Force component, but a cost reduction in the Navy component of the program (Hough, 1992). These changes must be tracked separately, but aggregated to acquire the complete picture of the program cost growth. Maritime acquisition provides a good example of a program that doesn't follow the typical milestone process; instead, the lead production contract serves a role similar to the EMD phase, and the follow-on production can be thought of as full rate production. These distinctions are not always clear in the SAR; some maritime acquisition efforts even report each hull number like a separate program (this is usually the case for aircraft carriers). For these reasons, each SAR must be examined carefully for unusual program divisions or departures from the typical acquisition profile.

### Conclusion

Significant work has been performed in the areas of SAR analysis, while macro-stochastic estimation techniques are still nascent. Best practices in all of these areas must be applied to the most up-to-date and salient dataset in order to produce the highest quality inferences and predictions. This chapter summarizes the process used in DoD cost estimation and affordability assessment, illustrating the utility of a tool for predicting changes in a program’s APB. This chapter also highlights commonly cited barriers to accurate SAR analyses. Some of these barriers are applicable to this research effort, but many are not, since this research does not seek to establish causation. The next chapter will discuss the aspects of former studies that are incorporated into this research, and list the relevant limitations and assumptions.

# III. Methodology

## Chapter Overview

This chapter explains filtering criteria used when collecting acquisition cost estimate data from Major Defense Acquisition Programs (MDAPs). It also discusses the statistical model, the model selection criteria, and the model selection technique. Finally, it presents the validation methodology for the predictive model.

## Dataset Formation

Submission of acquisition reports to Congress began in 1969 for select programs; therefore, the complete body of cost estimate information for these programs is vast. Unfortunately, constant changes to reporting requirements create dissimilarities that pose challenges to analysis. In order to restrict the dataset for this analysis, and ensure applicable and homogeneous data, five filtering criteria are applied to the available SAR data. It is difficult to determine the effect that these criteria will have on the performance of the final models *a priori*. For this reason, criteria are chosen that have been established by previous SAR analyses. In some cases, even more restrictive criteria are used to reduce the scope of this study to a manageable level. Chapter 5 assesses the impact of these criteria on the quality of the final model. These five filtering criteria are discussed below.

First, only MDAPs are considered. These programs historically comprise approximately 50 percent of the procurement budget (Jarvaise et al., 1996) and are required to report their status annually via the SAR. While other programs, such as Major Automated Information Systems (MAIS), report acquisition data to Congress, these programs are excluded from this study in order to make the scope more manageable.

The second major filtering criterion is the program initiation date. This study only includes programs with a Milestone B date of 1987 or later. Some SARs report planning SARs prior to this milestone, but the program is not considered an official Program of Record until Milestone B, and the program structure is not formally established until this milestone. The 1987 threshold serves several purposes. The year 1987 is the first year after the Packard Commission fundamentally reformed the DoD acquisition process. Therefore, this threshold prevents disparate reporting requirements and acquisition practices for older programs from biasing the results. Many of the other major revisions to SAR reporting requirements, occurred prior to 1987, allowing greater continuity in the dataset (Arena et al., 2006). Additionally, the selection of a threshold is necessary to scope the data reduction effort, and complete the study in the required time.

The third filtering criterion is the completion criterion. Since this study is primarily concerned with measuring acquisition cost estimate error over time, the completion criterion ensures that a program has reached a level of maturity sufficient to allow meaningful estimation of this error. However, requiring programs to have completed the entire acquisition phase is overly restrictive, due to the high average duration of these multi-billion dollar programs. Therefore, programs that have expended at least 50 percent of their projected funds, and have produced at least 25 percent of their planned units are included in the study. This completion threshold is more restrictive than those used in previous studies, as shown in Chapter 2. An exception is made to the 25 percent production requirement for Navy programs that procure maritime vessels.

These maritime acquisition programs will sometimes divide reporting into completion of individual vessels or lot-buys, showing no progress until all (or most) of the vessels are complete[1]. Allowing incomplete programs into the dataset necessarily sacrifices data quality for a sufficient sample size (both of which are required for a useful model). The impact of this criterion is assessed in Chapter 5. No cancelled programs are admitted to the dataset, though this criterion only omits one program.

Fourth, all programs must have at least four data points. Since MDAP status is based upon acquisition cost estimates, cost overruns may cause a program that is not initially designated as an MDAP to exceed the reporting requirement threshold with only a few years left in the acquisition phase. Such programs skew the results, as they meet the completion threshold, but do not have sufficient repeated measures to produce a meaningful estimation error. Therefore, only programs with four or more SARs are included in the study. This number is more restrictive than Drezner's threshold of three SARs (Drezner et al., 1993), ensuring that sufficient repeated measures are achieved to establish trends for each program.

Finally, this research allows changes to a program's baseline, but cannot utilize data for programs that are fundamentally restructured before completion[2]. Any large inconsistencies in ground rules and assumptions for generating estimates make it difficult

[1] CVN-68, AOE-6, MHC-51, and SSN-21 programs all go from 0% to over 80% acquired in a single SAR.

[2] For example, the Patriot Advanced Capability (PAC-3) program made four total changes to the structure of the program. Sometimes the baseline specified components of the system and only reported costs for select ones, while other times, the system was reported as a whole.

to accurately determine the cost estimate for a system. This criterion must be assessed on a case-by-case basis.

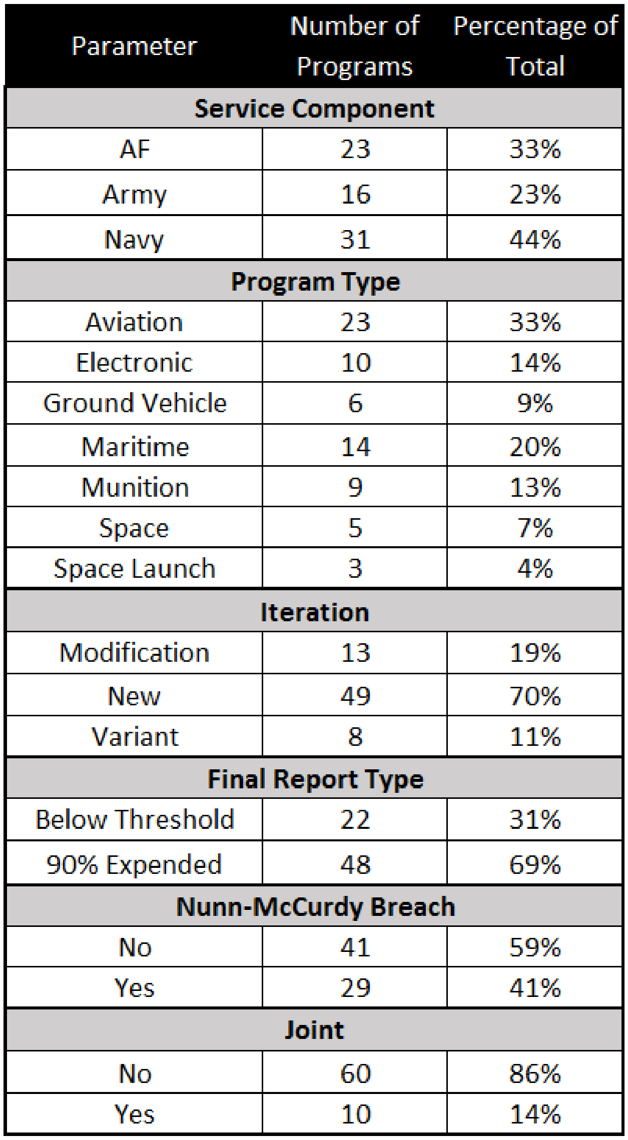
**Dataset Contents**

This section presents summary statistics for the final dataset. Out of over 319 MDAPs with initiation dates later than 1987, 70 programs (21 percent) qualify for entry into the dataset. There are an average of 13.4 SARs for each of these programs, resulting in a total of 937 program-years of data. The most recent SARs used are from 2012 (the most recent data available at the time of this study). Table 1 summarizes the data using different nominal parameters. Overall, the data provide sufficient observations in each category, which helps prevent divergent extrapolation—a condition where parameter combinations cause invalid inferences to be drawn where no collected data exists.

The *Service Component* variable in Table 1 indicates the DoD service component responsible for the program; in the case of a joint program, it indicates the lead service. The *Program Type* variable is based on the SAR *Mission and Description* section, as well as the appropriation category. These seven types are consistent with those used in the RAND analyses cited in Chapter 2.

The *iteration* variable indicates whether a program is a completely new system, a modification to an existing system, or a variant of an existing system. For example, the C-5 Avionics Modernization Program is a modification, but the F-18E/F program is a variant (it is a new version of an existing airframe). All systems with new letter designations (F-16C/D, F-14D) are considered variants.

**Table 1. Nominal Parameter Frequencies**



| Parameter | Number of Programs | Percentage of Total |
|---|---|---|
| **Service Component** | | |
| AF | 23 | 33% |
| Army | 16 | 23% |
| Navy | 31 | 44% |
| **Program Type** | | |
| Aviation | 23 | 33% |
| Electronic | 10 | 14% |
| Ground Vehicle | 6 | 9% |
| Maritime | 14 | 20% |
| Munition | 9 | 13% |
| Space | 5 | 7% |
| Space Launch | 3 | 4% |
| **Iteration** | | |
| Modification | 13 | 19% |
| New | 49 | 70% |
| Variant | 8 | 11% |
| **Final Report Type** | | |
| Below Threshold | 22 | 31% |
| 90% Expended | 48 | 69% |
| **Nunn-McCurdy Breach** | | |
| No | 41 | 59% |
| Yes | 29 | 41% |
| **Joint** | | |
| No | 60 | 86% |
| Yes | 10 | 14% |

The *Final Report Type* variable is useful for tracking the number of programs that are complete, versus those that are incomplete. The *Nunn-McCurdy Breach* parameter in Table 1 indicates whether a program has ever had such a breach. This breach, established in the 1982 Defense Authorization Act, is a formal measure of cost growth that requires an increased level of scrutiny from Congress. The Nunn-McCurdy breach is indicated as a binary variable in each SAR, located in the *Threshold Breaches* section.

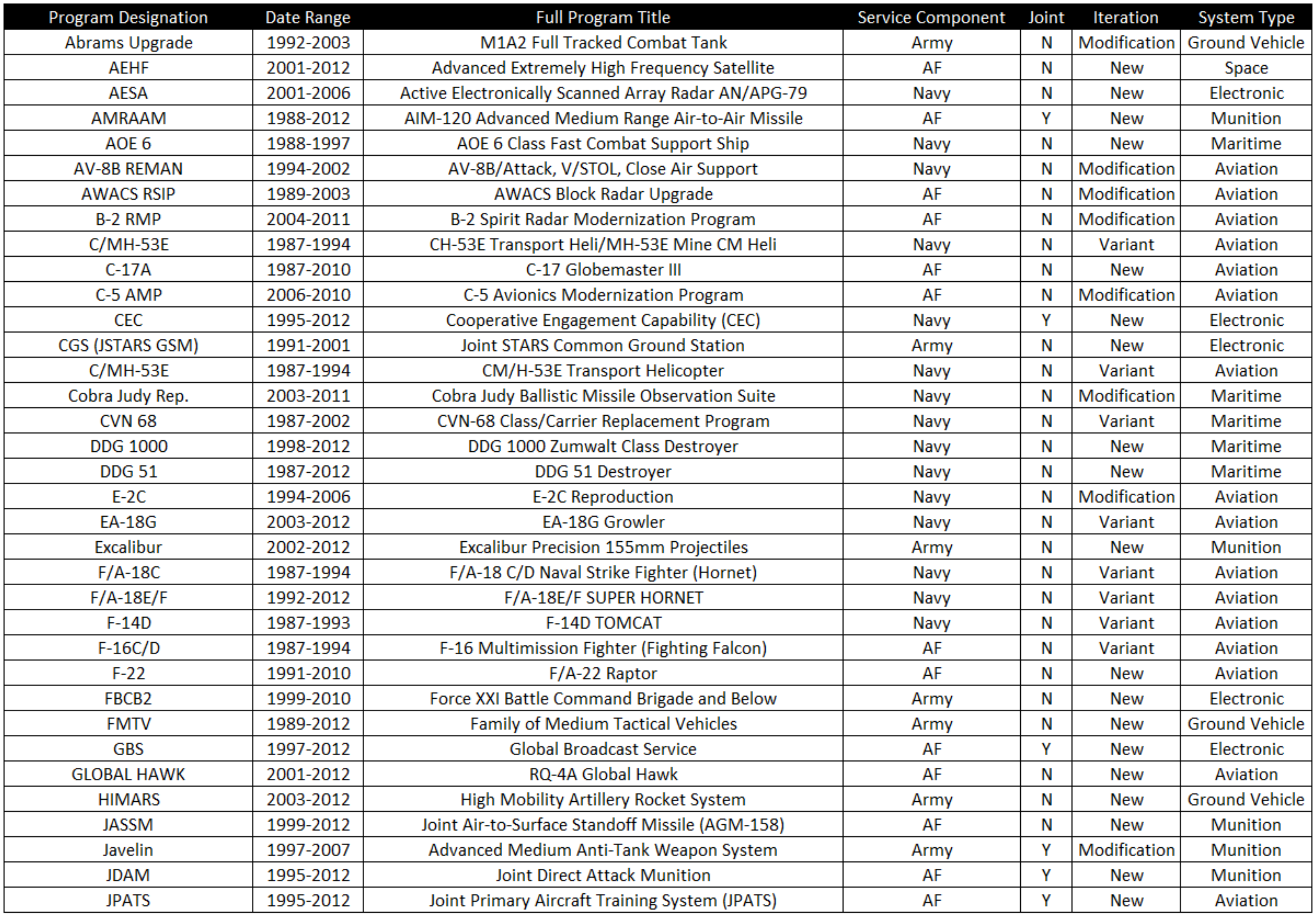
Programs that are jointly funded with another service are indicated in the SAR and none of the programs in this database switch their status during their acquisition phase, though this is a possibility with other MDAPs. Joint programs are identified in the SAR's *Program Information* section.

Table 2 lists the programs in the dataset, and provides summary-level descriptions of each. In addition to the levels of the parameters, described above, Table 2 also shows the span of program years, and the full title of the program.

### Data Verification

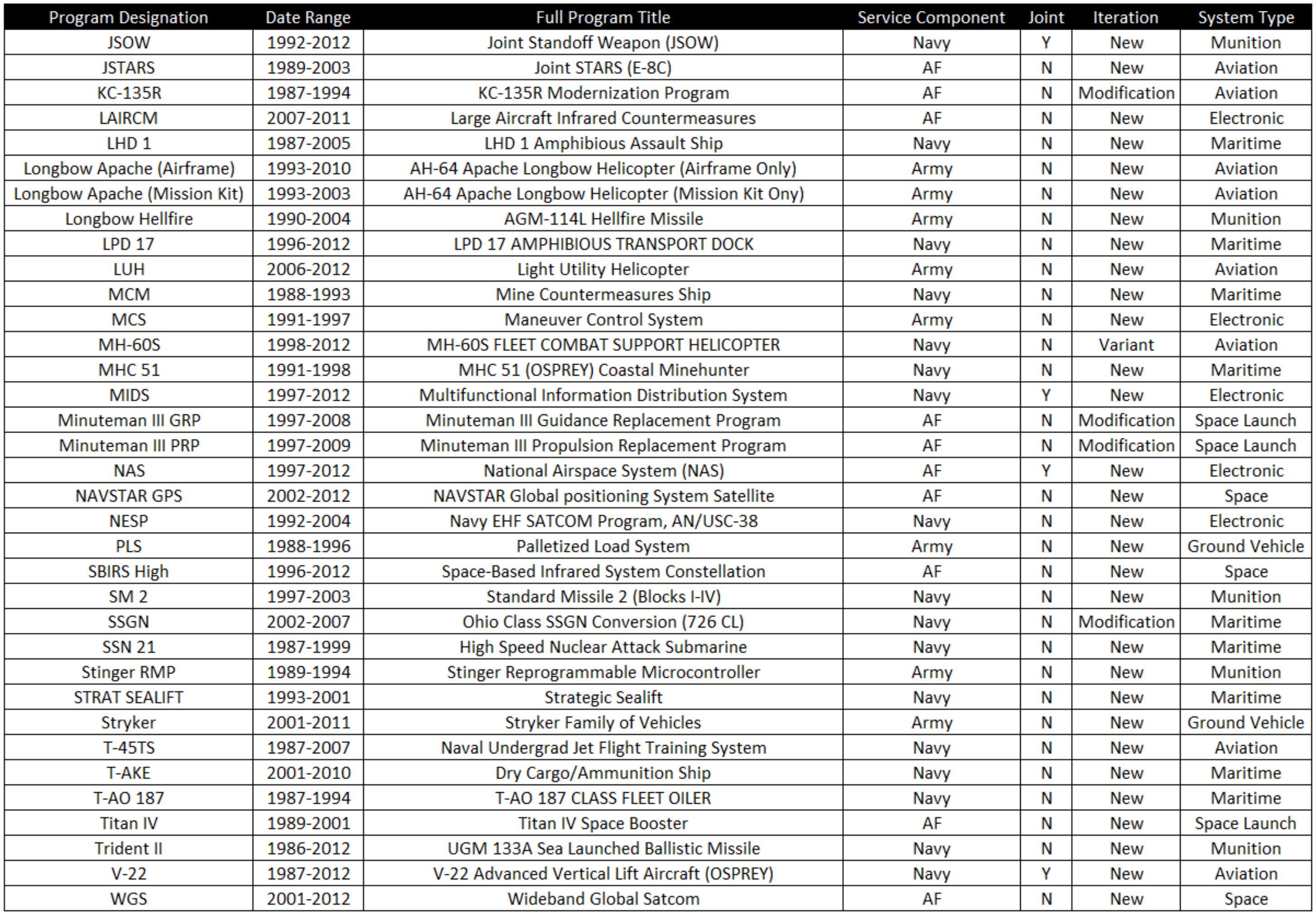
Prior to analysis, it is important to examine the data for any data entry or typographical errors. It is also useful to convert values to common units, so that transformations may be applied in a uniform manner. For example, variables with dollar amounts erroneously reported in units of thousands—a relatively common error—are corrected to be in millions. Dollar amounts *correctly* reported in thousands are also converted to millions in order to establish analytical continuity. This data verification and conversion process is performed as data is entered into the database. The distribution of each variable is also examined to identify outliers. However, these outlying observations are typically retained since the removal of data points that exhibit meaningful errors would adversely affect resource allocation. Outliers are only removed from this study if they can be attributed to typographical or data entry errors, and such errors cannot be corrected.

## Table 2. Program Data Summary



| Program Designation | Date Range | Full Program Title | Service Component | Joint | Iteration | System Type |
|---|---|---|---|---|---|---|
| Abrams Upgrade | 1992-2003 | M1A2 Full Tracked Combat Tank | Army | N | Modification | Ground Vehicle |
| AEHF | 2001-2012 | Advanced Extremely High Frequency Satellite | AF | N | New | Space |
| AESA | 2001-2006 | Active Electronically Scanned Array Radar AN/APG-79 | Navy | N | New | Electronic |
| AMRAAM | 1988-2012 | AIM-120 Advanced Medium Range Air-to-Air Missile | AF | Y | New | Munition |
| AOE 6 | 1988-1997 | AOE 6 Class Fast Combat Support Ship | Navy | N | New | Maritime |
| AV-8B REMAN | 1994-2002 | AV-8B/Attack, V/STOL, Close Air Support | Navy | N | Modification | Aviation |
| AWACS RSIP | 1989-2003 | AWACS Block Radar Upgrade | AF | N | Modification | Aviation |
| B-2 RMP | 2004-2011 | B-2 Spirit Radar Modernization Program | AF | N | Modification | Aviation |
| C/MH-53E | 1987-1994 | CH-53E Transport Heli/MH-53E Mine CM Heli | Navy | N | Variant | Aviation |
| C-17A | 1987-2010 | C-17 Globemaster III | AF | N | New | Aviation |
| C-5 AMP | 2006-2010 | C-5 Avionics Modernization Program | AF | N | Modification | Aviation |
| CEC | 1995-2012 | Cooperative Engagement Capability (CEC) | Navy | Y | New | Electronic |
| CGS (JSTARS GSM) | 1991-2001 | Joint STARS Common Ground Station | Army | N | New | Electronic |
| C/MH-53E | 1987-1994 | CM/H-53E Transport Helicopter | Navy | N | Variant | Aviation |
| Cobra Judy Rep. | 2003-2011 | Cobra Judy Ballistic Missile Observation Suite | Navy | N | Modification | Maritime |
| CVN 68 | 1987-2002 | CVN-68 Class/Carrier Replacement Program | Navy | N | Variant | Maritime |
| DDG 1000 | 1998-2012 | DDG 1000 Zumwalt Class Destroyer | Navy | N | New | Maritime |
| DDG 51 | 1987-2012 | DDG 51 Destroyer | Navy | N | New | Maritime |
| E-2C | 1994-2006 | E-2C Reproduction | Navy | N | Modification | Aviation |
| EA-18G | 2003-2012 | EA-18G Growler | Navy | N | Variant | Aviation |
| Excalibur | 2002-2012 | Excalibur Precision 155mm Projectiles | Army | N | New | Munition |
| F/A-18C | 1987-1994 | F/A-18 C/D Naval Strike Fighter (Hornet) | Navy | N | Variant | Aviation |
| F/A-18E/F | 1992-2012 | F/A-18E/F SUPER HORNET | Navy | N | Variant | Aviation |
| F-14D | 1987-1993 | F-14D TOMCAT | Navy | N | Variant | Aviation |
| F-16C/D | 1987-1994 | F-16 Multimission Fighter (Fighting Falcon) | AF | N | Variant | Aviation |
| F-22 | 1991-2010 | F/A-22 Raptor | AF | N | New | Aviation |
| FBCB2 | 1999-2010 | Force XXI Battle Command Brigade and Below | Army | N | New | Electronic |
| FMTV | 1989-2012 | Family of Medium Tactical Vehicles | Army | N | New | Ground Vehicle |
| GBS | 1997-2012 | Global Broadcast Service | AF | Y | New | Electronic |
| GLOBAL HAWK | 2001-2012 | RQ-4A Global Hawk | AF | N | New | Aviation |
| HIMARS | 2003-2012 | High Mobility Artillery Rocket System | Army | N | New | Ground Vehicle |
| JASSM | 1999-2012 | Joint Air-to-Surface Standoff Missile (AGM-158) | AF | N | New | Munition |
| Javelin | 1997-2007 | Advanced Medium Anti-Tank Weapon System | Army | Y | Modification | Munition |
| JDAM | 1995-2012 | Joint Direct Attack Munition | AF | Y | New | Munition |
| JPATS | 1995-2012 | Joint Primary Aircraft Training System (JPATS) | AF | Y | New | Aviation |

## Table 2. Program Data Summary (Continued)



| Program Designation | Date Range | Full Program Title | Service Component | Joint | Iteration | System Type |
|---|---|---|---|---|---|---|
| JSOW | 1992-2012 | Joint Standoff Weapon (JSOW) | Navy | Y | New | Munition |
| JSTARS | 1989-2003 | Joint STARS (E-8C) | AF | N | New | Aviation |
| KC-135R | 1987-1994 | KC-135R Modernization Program | AF | N | Modification | Aviation |
| LAIRCM | 2007-2011 | Large Aircraft Infrared Countermeasures | AF | N | New | Electronic |
| LHD 1 | 1987-2005 | LHD 1 Amphibious Assault Ship | Navy | N | New | Maritime |
| Longbow Apache (Airframe) | 1993-2010 | AH-64 Apache Longbow Helicopter (Airframe Only) | Army | N | New | Aviation |
| Longbow Apache (Mission Kit) | 1993-2003 | AH-64 Apache Longbow Helicopter (Mission Kit Ony) | Army | N | New | Aviation |
| Longbow Hellfire | 1990-2004 | AGM-114L Hellfire Missile | Army | N | New | Munition |
| LPD 17 | 1996-2012 | LPD 17 AMPHIBIOUS TRANSPORT DOCK | Navy | N | New | Maritime |
| LUH | 2006-2012 | Light Utility Helicopter | Army | N | New | Aviation |
| MCM | 1988-1993 | Mine Countermeasures Ship | Navy | N | New | Maritime |
| MCS | 1991-1997 | Maneuver Control System | Army | N | New | Electronic |
| MH-60S | 1998-2012 | MH-60S FLEET COMBAT SUPPORT HELICOPTER | Navy | N | Variant | Aviation |
| MHC 51 | 1991-1998 | MHC 51 (OSPREY) Coastal Minehunter | Navy | N | New | Maritime |
| MIDS | 1997-2012 | Multifunctional Information Distribution System | Navy | Y | New | Electronic |
| Minuteman III GRP | 1997-2008 | Minuteman III Guidance Replacement Program | AF | N | Modification | Space Launch |
| Minuteman III PRP | 1997-2009 | Minuteman III Propulsion Replacement Program | AF | N | Modification | Space Launch |
| NAS | 1997-2012 | National Airspace System (NAS) | AF | Y | New | Electronic |
| NAVSTAR GPS | 2002-2012 | NAVSTAR Global positioning System Satellite | AF | N | New | Space |
| NESP | 1992-2004 | Navy EHF SATCOM Program, AN/USC-38 | Navy | N | New | Electronic |
| PLS | 1988-1996 | Palletized Load System | Army | N | New | Ground Vehicle |
| SBIRS High | 1996-2012 | Space-Based Infrared System Constellation | AF | N | New | Space |
| SM 2 | 1997-2003 | Standard Missile 2 (Blocks I-IV) | Navy | N | New | Munition |
| SSGN | 2002-2007 | Ohio Class SSGN Conversion (726 CL) | Navy | N | Modification | Maritime |
| SSN 21 | 1987-1999 | High Speed Nuclear Attack Submarine | Navy | N | New | Maritime |
| Stinger RMP | 1989-1994 | Stinger Reprogrammable Microcontroller | Army | N | New | Munition |
| STRAT SEALIFT | 1993-2001 | Strategic Sealift | Navy | N | New | Maritime |
| Stryker | 2001-2011 | Stryker Family of Vehicles | Army | N | New | Ground Vehicle |
| T-45TS | 1987-2007 | Naval Undergrad Jet Flight Training System | Navy | N | New | Aviation |
| T-AKE | 2001-2010 | Dry Cargo/Ammunition Ship | Navy | N | New | Maritime |
| T-AO 187 | 1987-1994 | T-AO 187 CLASS FLEET OILER | Navy | N | New | Maritime |
| Titan IV | 1989-2001 | Titan IV Space Booster | AF | N | New | Space Launch |
| Trident II | 1986-2012 | UGM 133A Sea Launched Ballistic Missile | Navy | N | New | Munition |
| V-22 | 1987-2012 | V-22 Advanced Vertical Lift Aircraft (OSPREY) | Navy | Y | New | Aviation |
| WGS | 2001-2012 | Wideband Global Satcom | AF | N | New | Space |

The database contains two types of variables—or *factors*—extracted from data in each SAR: reported factors, and calculated factors.  Reported factors are read directly from the SAR text, and entered into the database as presented.  Examples of a reported factor are *Service Component* and *Iteration*.  The database contains thirty-nine reported factors.  Using combinations of the reported factors, mathematical operations are performed to generate forty-nine other variables, called calculated factors.  An example of a calculated factor is *Years since Milestone B*, which uses the *MS-B* and *SAR Date* reported factors to calculate the new variable.  All reported factors analyzed in this study are summarized in Table 3, and all calculated factors are shown in Table 4.  Many of these calculations are discussed in greater detail in the following sections.

The factor level in Table 3 indicates whether the value of a factor remains the same for the duration of the program (designated as *Program*), or may vary across SARs within a program (designated as *SAR*).  Factors that vary across SARs form the basis for the trajectory that a program's cost estimates take through the life of the program.  Some parameters, such as *Last Year of Production* have the SAR and Program level values recorded, since a given program will report the last year of production in each SAR, but a program has only one *true* last year of production. This true value is assumed to be the one reported in the final SAR for that program. The *SAR/Program* description in Table 3 indicates that both of these levels are retained.

## Table 3. Reported Factors

| Factor Name | Description | Level | Variable Type |
|---|---|---|---|
| **Program Name** | The formal name of the acquisition program | Program | Nominal |
| **Year** | The year of the SAR | SAR | Discrete |
| **SAR Date** | The date (day/month/year) of the SAR | SAR | Continuous |
| **Base Year** | The baseline year for cost reporting | SAR/Program | Discrete |
| **Service Component** | The lead service for the program | Program | Nominal: *AF*, *Navy*, or *Army* |
| **Joint** | Indicates that a program is funded by multiple services | Program | Binary |
| **Iteration** | Separates new programs from modifications and variants | Program | Nominal: *New*, *Modification*, or *Variant* |
| **Type** | Divides programs into mutually exclusive categories | Program | Nominal[4] |
| **Phase** | Indicates the program phase for each SAR | SAR | Nominal: *Development* or *Production* |
| **Final Report** | Indicates the status of the program for the final year of data | Program | Nominal: *Complete*, or *Below Threshold* |
| **Current APB** | Indicates the year of the Acquisition Program Baseline for each | SAR | Discrete |
| **Dev APB** | Indicates the year of the first development phase Acquisition | Program | Discrete |
| **Prod APB** | Indicates the year of the first Production Phase Acquisition | Program | Discrete |
| **RDT&E** | Research, Development, Test and Evaluation cost estimate (in | SAR | Continuous |
| **Procurement** | Procurement acquisition cost estimate (in SAR base year $) | SAR | Continuous |
| **MILCON** | Military Construction acquisition cost estimate (in SAR base year | SAR | Continuous |
| **Acq O&M** | Acquisition Operation and Maintenance cost estimate (in SAR base | SAR | Continuous |
| **Total ($M)**[1, 2, 3] | Reported total acquisition cost estimate (in SAR base year $) | SAR | Continuous |
| **Percent Expended** | Percent of program funds expended to date | SAR | Continuous |
| **Years Funded**[1, 2] | Number of years a program is funded from initiation | SAR/Program | Discrete |
| **APUC: Initial Dev Baseline** | The initial development baseline for Average Procurement Unit Cost | Program | Continuous |
| **APUC: Initial Prod Baseline** | The initial production basline for Average Procurement Unit Cost | Program | Continuous |
| **APUC: Current** | The current Average Procurement Unit Cost estimate | SAR | Continuous |
| **PAUC:Initial Dev Baseline** | The initial development baseline for Program Acquisition Unit Cost | Program | Continuous |
| **PAUC: Initial Prod Baseline** | The initial production baseline for Program Acquisistion Unit Cost | Program | Continuous |
| **PAUC: Current** | The current Program Acquisition Unit Cost estimate | SAR | Continuous |
| **EngrVar** | Cost variance due to engineering changes ($M in SAR base year) | SAR | Continuous |
| **EstVar** | Cost variance due to estimation assumption changes ($M in SAR base year) | SAR | Continuous |
| **QtyVar** | Cost variance due to Quantity changes ($M in SAR base year) | SAR | Continuous |
| **TotalVar** | Total cost variance from previous SAR | SAR | Continuous |
| **Schedule Breach** | Indicates that a program suffered a schedule breach | SAR | Binary |
| **Tech Perf Breach** | Indicates that a program suffered a technical performance breach | SAR | Binary |
| **Cost Breach** | Indicates that a program suffered a program cost breach | SAR | Binary |
| **PAUC/ APUC Breach** | Indicates that a program suffered a unit cost breach | SAR | Binary |
| **N/M** | Indicates if a program has ever experienced a Nunn-McCurdy Breach | Program | Binary |
| **MS-B** | The date of Milestone B (sometimes called Milestone II) | Program | Continuous |
| **Last Year of Production** | Indicates the last year of production as reported in the current SAR | SAR/Program | Discrete |
| **Original Quantity** | Production quantity from initial Acquisition Program Baseline | Program | Discrete |
| **Current Quantity** | Production quantity currently planned | SAR | Discrete |

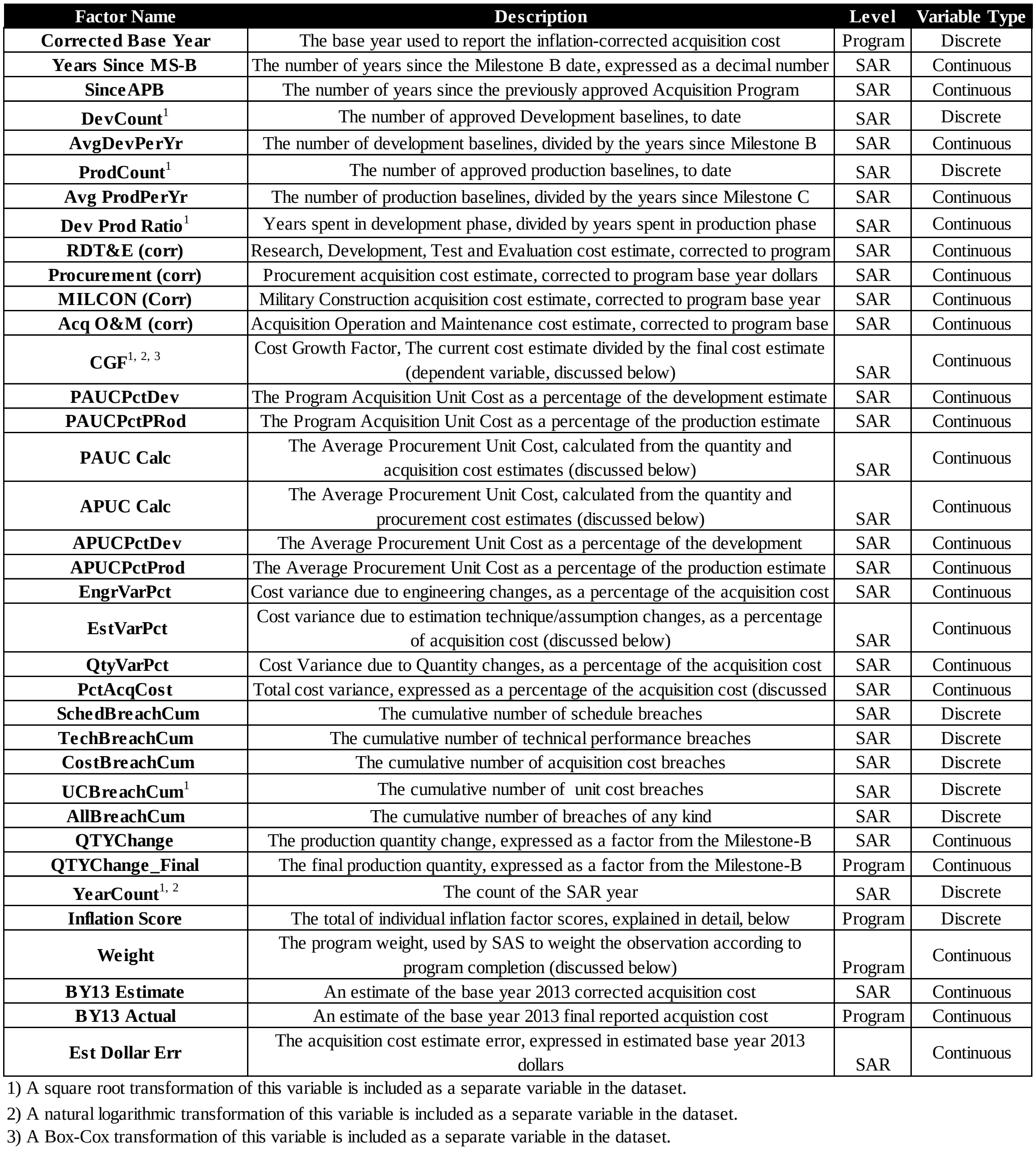
1) A natural logarithmic transformation of this variable is included as a separate variable in the

2) A square root transformation of this variable is included as a separate variable in the dataset.

3) A Box-Cox transformation of this variable is included as a separate variable in the dataset.

4) *Type* categories are: *Aviation, Electronic, Ground Vehicle, Maritime, Munition, Space, S*

## Table 4. Calculated Factors



| Factor Name | Description | Level | Variable Type |
|---|---|---|---|
| **Corrected Base Year** | The base year used to report the inflation-corrected acquisition cost | Program | Discrete |
| **Years Since MS-B** | The number of years since the Milestone B date, expressed as a decimal number | SAR | Continuous |
| **SinceAPB** | The number of years since the previously approved Acquisition Program | SAR | Continuous |
| **DevCount**[1] | The number of approved Development baselines, to date | SAR | Discrete |
| **AvgDevPerYr** | The number of development baselines, divided by the years since Milestone B | SAR | Continuous |
| **ProdCount**[1] | The number of approved production baselines, to date | SAR | Discrete |
| **Avg ProdPerYr** | The number of production baselines, divided by the years since Milestone C | SAR | Continuous |
| **Dev Prod Ratio**[1] | Years spent in development phase, divided by years spent in production phase | SAR | Continuous |
| **RDT&E (corr)** | Research, Development, Test and Evaluation cost estimate, corrected to program | SAR | Continuous |
| **Procurement (corr)** | Procurement acquisition cost estimate, corrected to program base year dollars | SAR | Continuous |
| **MILCON (Corr)** | Military Construction acquisition cost estimate, corrected to program base year | SAR | Continuous |
| **Acq O&M (corr)** | Acquisition Operation and Maintenance cost estimate, corrected to program base | SAR | Continuous |
| **CGF**[1, 2, 3] | Cost Growth Factor, The current cost estimate divided by the final cost estimate (dependent variable, discussed below) | SAR | Continuous |
| **PAUCPctDev** | The Program Acquisition Unit Cost as a percentage of the development estimate | SAR | Continuous |
| **PAUCPctPRod** | The Program Acquisition Unit Cost as a percentage of the production estimate | SAR | Continuous |
| **PAUC Calc** | The Average Procurement Unit Cost, calculated from the quantity and acquisition cost estimates (discussed below) | SAR | Continuous |
| **APUC Calc** | The Average Procurement Unit Cost, calculated from the quantity and procurement cost estimates (discussed below) | SAR | Continuous |
| **APUCPctDev** | The Average Procurement Unit Cost as a percentage of the development | SAR | Continuous |
| **APUCPctProd** | The Average Procurement Unit Cost as a percentage of the production estimate | SAR | Continuous |
| **EngrVarPct** | Cost variance due to engineering changes, as a percentage of the acquisition cost | SAR | Continuous |
| **EstVarPct** | Cost variance due to estimation technique/assumption changes, as a percentage of acquisition cost (discussed below) | SAR | Continuous |
| **QtyVarPct** | Cost Variance due to Quantity changes, as a percentage of the acquisition cost | SAR | Continuous |
| **PctAcqCost** | Total cost variance, expressed as a percentage of the acquisition cost (discussed | SAR | Continuous |
| **SchedBreachCum** | The cumulative number of schedule breaches | SAR | Discrete |
| **TechBreachCum** | The cumulative number of technical performance breaches | SAR | Discrete |
| **CostBreachCum** | The cumulative number of acquisition cost breaches | SAR | Discrete |
| **UCBreachCum**[1] | The cumulative number of unit cost breaches | SAR | Discrete |
| **AllBreachCum** | The cumulative number of breaches of any kind | SAR | Discrete |
| **QTYChange** | The production quantity change, expressed as a factor from the Milestone-B | SAR | Continuous |
| **QTYChange_Final** | The final production quantity, expressed as a factor from the Milestone-B | Program | Continuous |
| **YearCount**[1, 2] | The count of the SAR year | SAR | Discrete |
| **Inflation Score** | The total of individual inflation factor scores, explained in detail, below | Program | Discrete |
| **Weight** | The program weight, used by SAS to weight the observation according to program completion (discussed below) | Program | Continuous |
| **BY13 Estimate** | An estimate of the base year 2013 corrected acquisition cost | SAR | Continuous |
| **BY13 Actual** | An estimate of the base year 2013 final reported acquistion cost | Program | Continuous |
| **Est Dollar Err** | The acquisition cost estimate error, expressed in estimated base year 2013 dollars | SAR | Continuous |

1) A square root transformation of this variable is included as a separate variable in the dataset.

2) A natural logarithmic transformation of this variable is included as a separate variable in the dataset.

3) A Box-Cox transformation of this variable is included as a separate variable in the dataset.

Tables 3 and 4 also show what transformations, if any, have been applied to variables in an attempt to linearize them. Transformed variables have a mathematical operation, such as the square root, applied to every observation for that variable. The goal of this transformation is to linearize the observations so that they may be predicted by the linear model, and exhibit a normal distribution around the average value. Since these transformations are tracked as separate variables in the analysis, they are added to the total number of calculated parameters.

Many linearizing transformations can be accomplished by raising the variable to some exponent. For example, the square root transformation mentioned above is equivalent to raising the variable to the (1/2) power; an inverse transformation is equivalent to raising the variable to the (-1) power. A common transformation technique, called the Box-Cox transformation, uses a method in which the parameter of interest has a variable exponent ($\lambda$) placed on it, and this exponent is varied through a range of specified values to find the one that best transforms the variable so that it exhibits normally distributed residuals (Box and Cox, 1964). The chosen value of $\lambda$ is then rounded to the nearest common transform while maintaining the properties of the best transform. The best convenient value for $\lambda$ on the CGF parameter is $\lambda$= -0.3, which is approximately the inverse cube root. The only other Box-Cox transformed variable, as shown in Table 3, is the total acquisition cost. An inverse square root transformation is applied to this variable ($\lambda$= -0.5).

Calculation of new factors from reported factors is expected to induce collinearity with these recorded factors; however, when collinearity occurs, the best performing correlated parameter is retained in the final model. Discarding less useful parameters with high correlation will improve the model, since model selection uses a metric enforce parsimony, as described below.

### Data Normalization

As mentioned in Chapter 2, it is typical for SAR analyses to "maintain the integrity of the baseline" by normalizing cost estimates to control for changes in quantity and inflation (Drezner et al., 1993:11). For the purposes of this research, adjusting for changes in quantity would mask this especially large predictor of cost growth—recent studies attribute nearly 40 percent of cost growth to quantity changes alone (GAO, 2012a). For this reason, cost estimates are *not* adjusted for quantity.

Inflation, however, is a nuisance factor since it creates a significant trend in the data, but is not a parameter of interest in this analysis. Additionally, it affects most programs equally, and can disguise other sources of estimate error. Therefore, the data are corrected for inflation through the use of constant Base Year dollars. Correcting for inflation is often unnecessary as most SARs for a given program are already reported in constant Base Year dollars. This allows direct comparison of cost estimates within a given program. Unfortunately, programs may change their Base Year when a new acquisition program baseline (APB) is established (a Base Year change is common at the start of the production phase). When this occurs, the incongruous data must be corrected so that later estimates are directly comparable to the initial estimate.

To correct program acquisition costs from one Base Year to another, each of the four components of the total acquisition cost estimate must be recorded, since these components each have a unique inflation index. These component costs are:

- Research, development, test and evaluation (RDT&E)
- Procurement
- Military construction (MILCON)
- Acquisition phase Operating and Maintenance (Acq. O&M)

Each of these components has its own inflation index. The inflation rates for Navy and Army are published annually, and easily accessed using the Joint Inflation Calculator. The inflation rates for the Air Force are also published annually, and may be accessed in a variety of useful tools. For this study, the Air Force inflation rates are extracted from the Air Force Financial Management and Comptroller 2012 version of the Excel-based plugin that functions like a calculator (SAF/FMCE, 2012).

Performing Base Year corrections causes discontinuities in two other cost estimates. The first of these is the unit cost estimate. The expected unit cost for a program is estimated in the SAR using two metrics: The Average Procurement Unit Cost (APUC) and the Program Acquisition Unit Cost (PAUC). To maintain the link between the unit cost estimate and the program cost estimate, the APUC and PAUC are re-calculated from the corrected Base Year program estimate, as shown in equations 1 and 2, below.

$$APUC_{CALC} = \frac{Base\ Year\ Corrected\ Procurement\ Estimate}{Procurement\ Quantity} \tag{1}$$

$$PAUC_{CALC} = \frac{Base\ Year\ Corrected\ Program\ Cost\ Estimate}{Procurement\ Quantity} \tag{2}$$

This recalculation also overcomes the challenge of missing data, since unit cost estimates are not reported in Base Year dollars until several years after acquisition cost estimates begin reporting in a common Base Year. Therefore, wherever Base Year data are missing, or Base Year costs have been corrected to another Base Year, the calculated unit cost estimates are used.

Cost variance is the second discontinuity caused by correcting to a different Base Year. The equations for calculating cost variance are more complicated, and the parameters for these calculations are not recorded in the SAR. For example, cost variance due to economic considerations is “a change that is solely due to price-level changes in the economy” (Hough, 1992:5). The source data used to calculate economic variance are not given in the SAR. To maintain continuity, the original cost variance numbers are used to calculate the annual and cumulative percent change using the original Base Year. This normalization technique allows these numbers to remain applicable when changing from one Base Year to another.

This method of describing factors as a percentage of the total acquisition cost is useful for normalizing programs of different sizes and years. However, the impact of this research is best conveyed through the use of actual dollars, since that is the natural measure for cost estimates.  As mentioned previously, the four appropriation categories that make up the acquisition cost estimate may be used to correct all programs to the same Base Year in order to make a meaningful comparison.  These components were not part of the data collection effort for programs that have a single Base Year; therefore, this inflation factor is estimated to give an approximate idea of the model's efficacy in terms of Base Year 2013 dollars saved.  In order to report SAR estimates in terms of 2013 Base Year dollars, the inflation rate for the appropriation categories is estimated.  Very few of the programs in this dataset have any MILCON or Acquisition O&M funding; when they do, the amount is typically less than ten percent of the overall amount.  Therefore, the separate inflation rates for these two cost components are disregarded in cases where the MILCON and Acquisition O&M amounts are unknown.  For the two remaining factors—Procurement and RDT&E funds—the raw inflation rates are averaged.  This average inflation rate differs by less than 1 percent from the actual inflation rate in any category, for any of the years in the applicable date range.  For the Army, the average rate is equal to the funding-specific rates, since the same inflation rate is used for all of the applicable categories.  The *BY13 Estimate*, and *BY13 Actual* variables multiply the total acquisition cost by the appropriate averaged rate to correct the reported dollar amounts into an estimated 2013 Base Year dollar amount.  Because of the error involved with the average inflation rate, this estimate is not used in any regression or model-building activities; it is only used to estimate the impact of the results in terms of real dollars.

### Dependent Variable

Since this study is concerned with improving the accuracy of early acquisition cost estimates, the parameter of interest is the error in these estimates. The accuracy of each estimate is expressed as a ratio between the current cost estimate and the actual program cost. This ratio, defined as the Cost Growth Factor (CGF), is calculated for each SAR. For example, consider an acquisition program with ten SARs. For the purposes of this research, the tenth and final SAR establishes the final program cost. The nine previous estimates are likely to differ from the actual program cost, overestimating or underestimating the final cost by varying degrees. This relationship is shown in equation 3, below.

$$CGF_i = \frac{\text{Actual Program Cost}}{\text{Cost Estimate } i} \tag{3}$$

Where $i$ is the number of the cost estimate, numbered sequentially from the initial estimate starting at 1

Overestimates (that is, coming in under budget) are illustrated by CGF values less than one, and underestimates (that is, cost overruns) are illustrated by CGF values greater than one. Intuitively, the cost estimate error for a given SAR may be calculated by taking one minus the CGF. Once a program's predicted CGF is calculated for the initial estimate—that is, the first SAR after Milestone B—then this estimate can be corrected to equal the actual program cost by simply multiplying the estimate by the predicted CGF.

One clarification must be made to the definition of CGF provided above: the actual cost of a program is not explicitly specified in the SAR. For the purposes of this research, it is assumed that the final estimate is the actual program cost, although SARs are only required until a program is 90 percent expended, or has delivered 90 percent of its units.  In the cases where the program is at least 92.5 percent expended, this estimate is shown not to be statistically different from final program cost (Tracy, 2005).  Some of the estimates in the dataset meet or exceed this 92.5 percent threshold, though the vast majority do not.  The implications of this assumption are discussed in Chapter 5.

The practice of calculating the CGF for each program year, rather than just the initial year, allows the construction of trajectories that aid the predictive capability of the model.  As a program progresses, it may exhibit significant patterns in certain predictors that affect the estimate error in a predictable way.  For example, the procurement quantity may not be a significant predictor of CGF.  However, a *change* in the procurement quantity may be associated much more strongly with CGF.  Therefore, this analysis methodology embraces the longitudinal nature of the SAR data in order to draw inferences.

### Statistical Model

Longitudinal data are characterized as an aggregation of measurements taken on the same subject over time. These repeated measurements across time violate the assumption of independence that is common in general linear models.  Furthermore, non-uniform reporting intervals and missing data further violate assumptions made by these general models.  Features of this dataset include:

- Dependence of data across time,
- Non-constant variance,
- Missing observations, and
- Non-uniform measurements.

While this study assumes independence between programs, it must account for the lack of independence within programs. For example, the correlation between two consecutive cost estimates (say, the 2004 and 2005 F-22 cost estimates) is expected to be higher on average than the correlation between estimates from two programs (for example, the 2004 F-22 estimate and the 2004 F-16 estimate ). Therefore, the correlation between programs is assumed to be zero, but the correlation within programs cannot be assumed to be zero. If unaccounted for, this dependence incorrectly inflates the variance, possibly resulting in a model that contains insignificant parameters (Patetta, 2002).

In addition to this dependence, observations between programs are not expected to be identically distributed, either. This assumption of identical distribution implies that the errors (and thus the response) of a given program exhibit similar variance, a common assumption in studies where measurements are taken from similar processes that result in similar variance. However, since novelty is an intrinsic trait of DoD acquisition programs, they are expected to exhibit disparate variance. In fact, if all programs exhibited similar variance, this study would not be necessary, since it would be a simple matter to calculate a prediction interval that enclosed some known percentage of the acquisition portfolio.

Finally, the SAR dataset is comprised of longitudinal data with missing observations and non-uniform measurement periods. For example, as noted earlier, very few programs delivered a SAR in 2000, or in 2008, due to the delay in the President's budget submission. While most programs submit SARs on an annual cycle, some programs experience unacceptable threshold breaches and are required to generate an out-of-cycle SAR. Missing data and such aperiodic measurements can cause errors in parameter estimate calculations unless allowances are made in the linear model structure to account for these inconsistencies.

To overcome these difficulties, a mixed-model approach is adopted. The mixed model is a more flexible formulation of the general linear model that adds random effect parameters to allow for differences between subjects, and also allows for "a more flexible specification of the covariance structure of the random errors" (Patetta, 2002:61). Both of these additions are useful. The random effect parameters allow for proper treatment of continuous data that do not follow levels prescribed by an experimental design, but are observed randomly. The flexible covariance matrix allows for the treatment of time-series dependence and non-uniform data through the introduction of additional model parameters. The mixed model takes the form shown in equation 4.

$$y = X\beta + Z\gamma + \varepsilon \tag{4}$$

Where:
y is the vector of observed responses
X is the design matrix of fixed predictor variables
$\beta$ is the vector of regression parameters (population-specific)
Z is the design matrix of random predictor variables
$\gamma$ is the vector of random-effect parameters (subject-specific)
$\varepsilon$ is the vector of random errors

The vector of regression parameters (*β*) contains the parameters that describe the whole population, and are assumed to result from fixed variables. For example, categorical descriptors such as Service Component and Program Type are fixed throughout the life of a program, and represent a mutually exclusive and collectively exhaustive set of factors for this data set. The vector of random-effect parameters (γ) contains all of the parameters that vary within a program. These random-effect parameters can account for program-specific deviations from the average profile. Isolating the fixed and random effects into two categories prevents the heterogeneity within programs from obfuscating the difference between programs. In other words, it is capable of accounting for variability that exists within a program that would otherwise be labeled a source of error in a general linear model.

The variance of the general linear model is said to take the form $\sigma^2\mathbf{I}$. That is, the only source of variance arises from the random errors, and these are assumed to be independent (between measurements) with constant variance. However, as discussed above, the variance of the Linear Mixed Model (LMM) can take on a different forms to account for the lack of these simplifying assumptions. The variance structure in the LMM takes the form in equation 5, where G and R are the two components of the variance structure, and are uncorrelated with each other (Kincaid, 2005).

$$\mathrm{Var}(\boldsymbol{Y}) = var\begin{bmatrix}\boldsymbol{\Upsilon}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\boldsymbol{G} & 0\\ 0 & \boldsymbol{R}\end{bmatrix} \qquad (5)$$

The **R** matrix represents the variances and covariances associates with the error terms of the model, and the **G** matrix, represents the variances and covariances associated with the random effects. This covariance structure must be modeled, and—depending on the selected structure—may contain relatively few, or a great many, parameters to estimate. Fortunately, it is only necessary to model one component of the variance.

Selection of a structure for the covariance matrix is the subject of many scholarly articles, and no best method has been established. One author states that "One important question which, unfortunately, still has no good answer is how to select the covariance structure" (Kincaid, 2005:1). The initial data exploration and model construction allows estimation of different covariance structures that are used during selection of the final statistical model. Four covariance structures are assessed during the model-building phase: first-order autoregressive, compound symmetry, Toeplitz, and unstructured. The elements of the covariance matrix under each of these structures are summarized in Table 5. The first-order autoregressive—abbreviated as *AR(1)*—structure assumes that consecutive observations on the same subject are correlated, but this correlation decreases by a factor ($\rho$) as the distance between observations increases. The first-order autoregressive structure is expected to be the most appropriate for the data, since most consecutive observations are correlated by some amount, and this correlation is expected to decrease with successive estimates. The compound symmetry structure assumes a correlation between measurements (variances, on the diagonal) but assumes that all of the off-diagonal covariances are homogenous, regardless of proximity. The Toeplitz structure is a more general case of the AR(1) structure, which assumes correlation based on proximity, but allows the correlation to follow different patterns. Finally, the unstructured covariance, as the name suggests, allows every variance and covariance to be modeled. Other structures may be specified in the statistical software, but only these four are considered to scope this design effort.

**Table 5. Selected Covariance Structures**

| Structure | ($i$,$j$)th element |
|---|---|
| **Autoregressive (1)** | $\sigma^2\rho^{\|i-j\|}$ |
| **Compound Symmetry** | $\sigma_1 + \sigma^2\mathrm{I}(i = j)$ |
| **Toeplitz** | $\sigma_{\|i-j\|+1}$ |
| **Unstructured** | $\sigma_{ij}$ |

To estimate the efficiency of both the covariance structure and the model parameters, a range of candidate model parameters are tested for significance along with a range of candidate covariance structures. Then a model selection criterion is used to determine if the model is better or worse than the previous model. The model selection criterion used is the Bayesian Information Criterion (BIC). The BIC is a member of a family of similar “information criteria” that penalize overly-large models (Kutner et al., 2004:359). This penalty is required since the addition of predictors to a model will almost always increase the accuracy of the model, but such models quickly become very cumbersome and may overfit the observed data. The model selection criterion computes an efficiency factor that is increased as the model approximation gets better, but is penalized as more parameters are added. This method ensures a *parsimonious* model by selecting the smallest number of parameters that best describe the data.

## Model Selection

Many modern statistical software packages have routines that automatically select the most parsimonious model from user-selected criteria, such as BIC. However, no such automatic selection procedure exists for mixed models. Also, simply running every possible combination of parameters and performing the BIC calculation is unwieldy, since the 75 variables shown in Tables 3 and 4 may be combined to form 3.77x$10^{22}$ main effect combinations—that is, 0.38 sextillion[4]. Since the mixed model may include main effects, multi-factor interactions, and random effects (along with their multi-factor interactions), the actual number of feasible combinations is much higher. Clearly, this many combinations is prohibitively large, even for a computer.

Model selection is performed by manually testing combinations of parameters using statistical software and observing their effect on the BIC in an iterative fashion. Testing single parameters for significance individually is informative but insufficient, since multicollinearity and conditional significance may cause one previously significant parameter to become insignificant while unnecessarily inflating the variance in the model. Thousands of combinations are examined and the resulting predictive capability is periodically tested.

[4] This is calculated by computing the feasible subsets of model combinations: $\binom{75}{75} + \binom{75}{74} + \cdots + \binom{75}{1}$

This model selection procedure is performed in SAS version 9.3 using the *mixed* procedure. This procedure allows the specification of fixed effect terms, random effect terms, and a covariance structure, among other modeling criteria. SAS automatically computes the BIC for each model and outputs statistics used to assess the model validity, such as residual plots, and normal quantile plots. These outputs are discussed in detail in Chapter 4, and the SAS code used in this analysis is provided in Appendix A.

As discussed above, the filtering criteria for database formation require that programs are 50 percent expended and 25 percent procured. This threshold allows 22 of the 70 programs into the dataset that may be deemed relatively mature, but not complete. Since these programs may still have substantial program life that could alter predictions of the CGF, and since the final year of the acquisition phase is uncertain, these observations are weighted to reduce their effect on the model. The weighting scheme for each program is determined by the final estimate for that program, using the relationship in equation 6.

$$\text{Weight}_p = \begin{cases} 1, & \text{For completed programs} \\ P_e, & \text{Otherwise} \end{cases} \qquad (6)$$

Where $P_e \equiv$ Percentage of program acquisition cost expended to date

In equation 6, $P_e$ is defined as the percentage of a program's funding that has been expended at the time of the SAR. However, since the program funding is subject to change, the percentage of these funds that is expended will sometimes behave paradoxically, seeming to decrease as a program is expanded. For the purposes of the weighting algorithm, every observation within a program is weighted according to the final estimate of the available funding and, therefore, the final value of $P_e$. For programs that are incomplete, the program funding may change in future baselines, invalidating the assumed model weight for that program.

### Descriptive versus Predictive Model

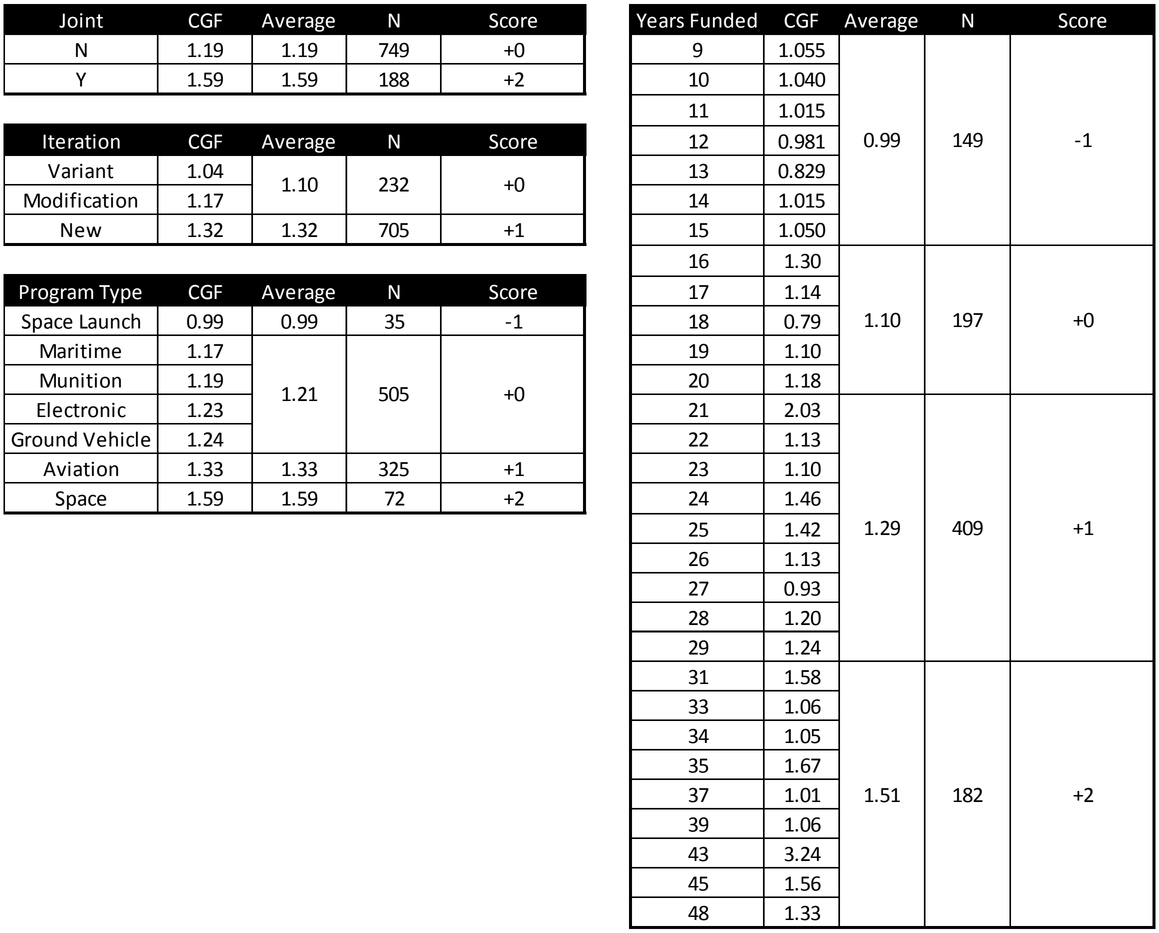
The first model created to describe CGF is the descriptive model. In this model, the individual acquisition programs are specified as the subjects, allowing the model to fit each program individually. This method results in 70 different regression models, and this level of specificity ensures a high degree of accuracy in predicting the past performance of the programs in the dataset. While this method may be used to demonstrate the validity of the macro-stochastic concept, it is not useful for predicting estimate error in future programs since the exact trajectory of these programs will not be repeated. Therefore, a predictive model must be developed that serves this application.

While the descriptive model specifies each program as its own subject, the opposite extreme—placing all programs into a single group—is not useful either. Since all programs are not expected to follow the same trajectory, this simplifying assumption decreases the resolution of the resulting model. The optimum predictive capability for the regression is achieved when subsets of programs are binned into some number of smaller groups. Then, the best model for predicting the trajectory of a new program may be applied by comparing the new program to the characteristics of the grouped programs.

The predictive model is formed under the assumption that future acquisition programs will follow the pattern of similar, past acquisition programs. This assumption implicitly requires a basis for comparison of similar programs (so that the appropriate group may be used to predict a new observation). For the purposes of this study, similarity is determined by the nominal program descriptors associated with trends in the CGF. For example, perhaps programs that have fixed macro-level descriptors associated with high cost growth will perform similarly to future programs with the same descriptors. These descriptors will need to contain fixed levels, since the variable levels of a new program must be known with certainty at program initiation.

To properly place programs into these categories, the cost growth must first be associated with different nominal variables. The four variables that are the most strongly associated with a trend in cost growth are selected, and each of these variables is split into levels. Table 6 shows that the four parameters used to score each program are: *Joint*, *Iteration*, *Program Type*, and *Years Funded*. Table 4 indicates the categorical levels associated with each parameter. For example, *Iteration* has three levels: *New*, *Modification* and *Variant*. However, these levels do not necessarily align with significant differences in cost growth (for example, the difference in cost growth between *Type* levels *Maritime* and *Munition* is small). Using every possible factor-level combination would produce too many program groups with too few programs in each group. Therefore, the levels of each variable are combined into groups that ensure a large sample size in the final program groups. Based on this balance, factor levels with little difference in their average CGF are combined, and the levels of the variables are determined as shown in Table 6.

### Table 6. Cost Growth Factor Contributors and Levels



| Joint | CGF | Average | N | Score |
|---|---|---|---|---|
| N | 1.19 | 1.19 | 749 | +0 |
| Y | 1.59 | 1.59 | 188 | +2 |

| Iteration | CGF | Average | N | Score |
|---|---|---|---|---|
| Variant | 1.04 | 1.10 | 232 | +0 |
| Modification | 1.17 | | | |
| New | 1.32 | 1.32 | 705 | +1 |

| Program Type | CGF | Average | N | Score |
|---|---|---|---|---|
| Space Launch | 0.99 | 0.99 | 35 | -1 |
| Maritime | 1.17 | 1.21 | 505 | +0 |
| Munition | 1.19 | | | |
| Electronic | 1.23 | | | |
| Ground Vehicle | 1.24 | | | |
| Aviation | 1.33 | 1.33 | 325 | +1 |
| Space | 1.59 | 1.59 | 72 | +2 |

| Years Funded | CGF | Average | N | Score |
|---|---|---|---|---|
| 9 | 1.055 | 0.99 | 149 | -1 |
| 10 | 1.040 | | | |
| 11 | 1.015 | | | |
| 12 | 0.981 | | | |
| 13 | 0.829 | | | |
| 14 | 1.015 | | | |
| 15 | 1.050 | | | |
| 16 | 1.30 | 1.10 | 197 | +0 |
| 17 | 1.14 | | | |
| 18 | 0.79 | | | |
| 19 | 1.10 | | | |
| 20 | 1.18 | | | |
| 21 | 2.03 | 1.29 | 409 | +1 |
| 22 | 1.13 | | | |
| 23 | 1.10 | | | |
| 24 | 1.46 | | | |
| 25 | 1.42 | | | |
| 26 | 1.13 | | | |
| 27 | 0.93 | | | |
| 28 | 1.20 | | | |
| 29 | 1.24 | | | |
| 31 | 1.58 | 1.51 | 182 | +2 |
| 33 | 1.06 | | | |
| 34 | 1.05 | | | |
| 35 | 1.67 | | | |
| 37 | 1.01 | | | |
| 39 | 1.06 | | | |
| 43 | 3.24 | | | |
| 45 | 1.56 | | | |
| 48 | 1.33 | | | |

Once the variable levels are established, they are combined for assigning a new program to a single group. To accomplish this, each variable level is assigned a score based upon its contribution to the CGF. For example, since the first level of the variable *Program Type* has a CGF of 0.99 on average, this level is assigned a score of -1, since it contributes a decrease in the CGF, on average. The average CGF in each variable level is used to assign such a score, according to equation 7, below. Then the total program score is used to bin programs according to a linear combination of their cost growth scores. The sum of these scores is called the *Cost Growth Score*.

$$\text{Cost Growth Score} = \begin{cases} -1 & \text{CGF} < 1 \\ 0 & 1 < \text{CGF} < 1.25 \\ +1 & 1.25 < \text{CGF} < 1.5 \\ +2 & \text{CGF} > 1.5 \end{cases} \tag{7}$$

With these scores established, each program in the dataset may be scored according to the observed levels of these four parameters. For example, the F-22 program is a new, non-joint aviation program that is funded for 34 years. Using Table 6 as a key, we see that this earns this program a Cost Growth Score of 4. By this method, each program is assigned a Cost Growth Score, and the resulting scores form the distribution shown in Figure 1. These six cost growth score bins are used as the subject in the predictive model. Note that bin six only contains three programs. This bin may have insufficient sample size for accurate predictions, a concern that is tested in the model validation step. Since the algorithm for grouping programs considers variables associated with cost growth, programs in the lower cost growth groups can be thought of as *low-growth* programs, while programs in the higher cost growth groups can be thought of as *high-growth* programs.

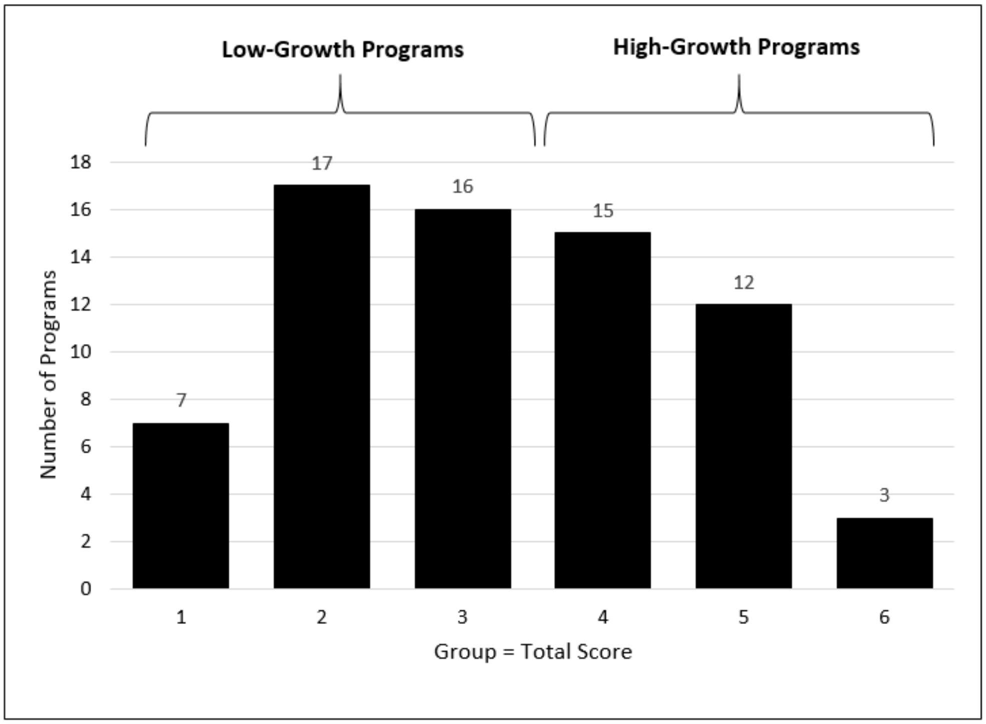


**Figure 1. Histogram of Program Cost Growth Group**

## Predictive Model Validation

Validation of the final predictive model demonstrates the ability to predict the acquisition cost estimate error of certain programs.  Since the data only sparsely populate certain factor-level combinations, it is undesirable to divide the data into a training and test data set for validation. Therefore, the entire dataset is used for model construction, and a modified version of the Leave One Out Cross Validation (LOOCV) method is used to validate the model.

The traditional LOOCV method involves fitting the model with all of the data, minus one observation, and then to assess the model's ability to accurately predict the dependent variable associated with this omitted observation. The traditional method is not sufficient in this case since the model is not trying to predict a single observation, but rather the entire trajectory of a new program. Therefore, this research requires that, instead of a single observation, an entire program is omitted, and the remaining programs in that category are used to predict the estimate error in the omitted program (Ryan et al., 2013). The omitted program is then incorporated into the data once again, and the next program is omitted. Then a new set of values for each model parameter are calculated and the CGF is predicted for each SAR for that program. In this way, an entire program becomes the observation. While the significance of model variables is determined with all data in the model, this technique ensures that the specific parameter estimates for each variable are determined without the knowledge of the program they are used to predict. In this way, the prediction capability of the model relative to each program category is shown by its ability to predict this omitted program. If poorly predicted programs exhibit a pattern (for example, if they are all Army munition programs), then this fact may be used to invalidate the model for that specific combination of parameters.

# IV. Analysis and Results

## Chapter Overview

In this chapter, the methodologies described in Chapter 3 are employed to conduct analyses and construct a descriptive and predictive model. The predictive model is validated using the modified Leave-One-Out Cross Validation (LOOCV) described in Chapter 3, and the resulting model efficacy is demonstrated.

## Uncorrected Error

The CGF for each estimate in the dataset may be examined to determine the average estimate error. This error, before any model corrections are applied, is referred to as "uncorrected" error in the discussion below. Figure 2 shows the error present in the SAR estimates plotted against program expenditure. The uncorrected CGF, averaged across programs as well as time, is 1.27, indicating that the average SAR from any program in any year is underestimating the actual program cost by 27 percent. However, this figure also confirms what we would intuitively expect: that program cost estimates are the least accurate near program initiation and improve with program maturity.

The CGF based on the first estimate of the program follows the distribution shown in Figure 3. This figure indicates that eleven programs in the data set (nearly 16 percent of the total) reported a final acquisition cost that exceeded their initial estimate by over 100 percent (a CGF greater than 2.0). In fact, only 25 of the 70 programs reported a final cost within 25 percent of their initial estimate. The mean value of the CGF from the first estimate is approximately 1.44, indicating that new MDAPs underestimate their

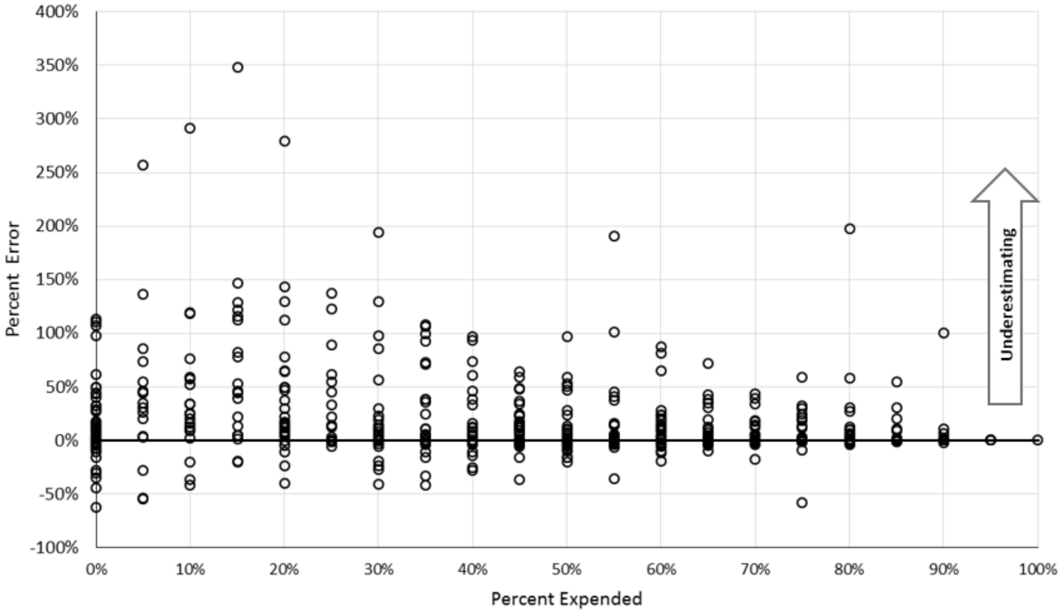


**Figure 2. Uncorrected Estimate Error by Percent Expended**

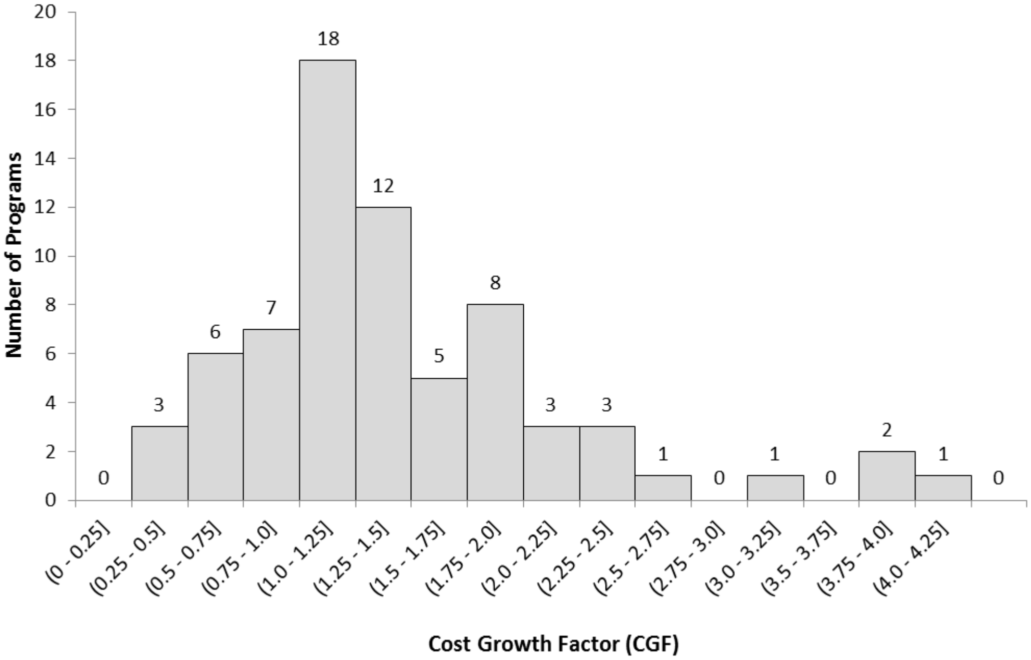
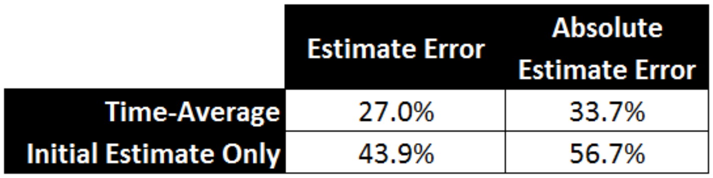


**Figure 3. Histogram of Acquisition Cost Growth Factor from Program Year One**

eventual cost by 44 percent, on average. Using only this mean value, we might consider simply adding 44 percent to the first cost estimate of all new MDAPs. However, the cost estimate error varies widely by program—the standard deviation from the average is 72 percent. Adding 44 percent to all initial program estimates might bring the average error closer to zero, but would not address this variation. In other words, the absolute deviation from the initial estimate is reduced by correcting programs individually (this is what the descriptive model does) or in groups (this is what the predictive model does).

Since overestimation and underestimation are considered equally detrimental for the purposes of this research, the absolute value of the estimate error provides additional insight when describing estimate error. The absolute value of the uncorrected estimate error is 34 percent, averaged across all programs and across time. Again, this error is worst at the outset of the program, with an absolute estimate error of 57 percent on average for the first estimate. This means that the average MDAP will have an eventual cost that is 57 percent different from what the initial estimate predicts. It is this initial, absolute error that the descriptive and predictive models are employed to reduce. The uncorrected error summary is presented by Table 7.

**Table 7. Average Uncorrected Cost Estimate Error**



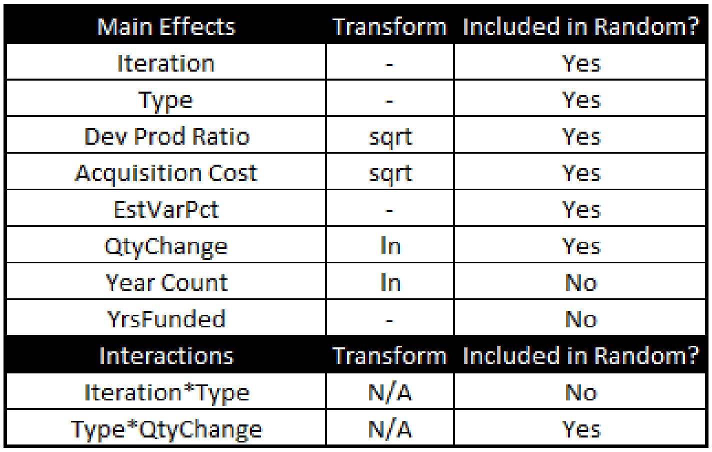
| | Estimate Error | Absolute Estimate Error |
|---|---|---|
| Time-Average | 27.0% | 33.7% |
| Initial Estimate Only | 43.9% | 56.7% |

## Descriptive Model

The results of the models take the form of predicted values for CGF. These predicted CGFs may be used to correct the observed error in each SAR, bringing the estimates more in line with the actual, eventual program cost. Recall that the descriptive model is formed by placing each program into its own category, allowing the regression to uniquely fit the model parameters to each program individually. Including the intercept, the descriptive model has eight main effects and eighteen terms in total. These variables are shown in Table 8. Note that the transformations performed on the interactions are the same transformations performed on the main effects; therefore, these are labeled “N/A.” For example, since the main effect for the *Quantity Change* variable was transformed using the natural logarithm, this same transform is used in the interaction. Variables that are in the main effects, but also included in the random effects, are indicated by a “Yes” in the column labeled “Included in Random?” Table 8 indicates that all parameters except *Year Count* and *Years Funded* are included as random effects. For the descriptive model SAS code, see Appendix A.

A first-order autoregressive—AR(1)—covariance structure best models the dependence within the cost estimate data. This structure resulted in a lower BIC for every examined combination of parameters, though the difference varied depending on the specific model being tested. The AR(1) covariance structure assumes that sequential observations are correlated; the “sequential correlation” parameter is a measure of how strongly these observations are related. The parameter estimate values for the sequential correlation ($\rho$) is estimated and shown with the model outputs in Appendix A.

**Table 8. Parameters in the Descriptive Model**

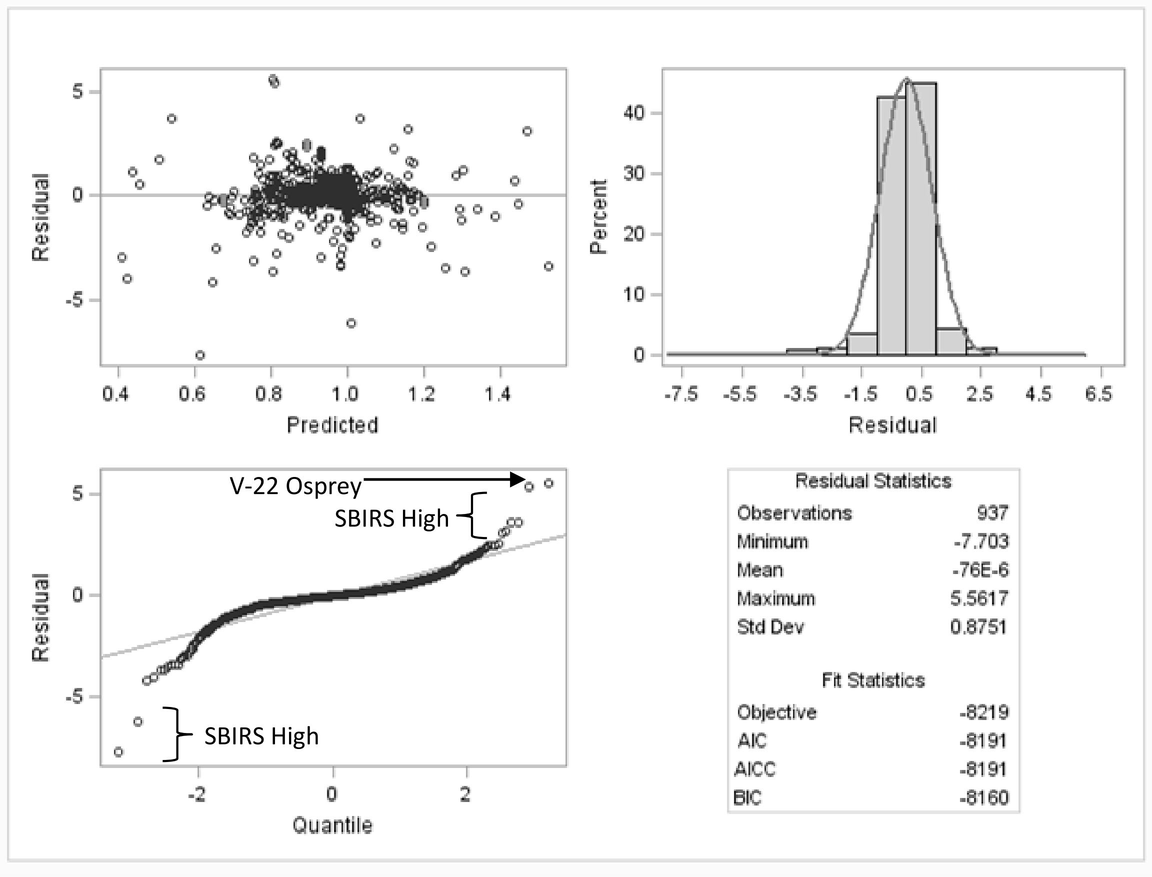


| Main Effects | Transform | Included in Random? |
|---|---|---|
| Iteration | - | Yes |
| Type | - | Yes |
| Dev Prod Ratio | sqrt | Yes |
| Acquisition Cost | sqrt | Yes |
| EstVarPct | - | Yes |
| QtyChange | ln | Yes |
| Year Count | ln | No |
| YrsFunded | - | No |
| **Interactions** | **Transform** | **Included in Random?** |
| Iteration*Type | N/A | No |
| Type*QtyChange | N/A | Yes |

## Descriptive Model Adequacy

Robust statistical models require that any variables included in the random terms must be included in the fixed effects as well, to avoid introducing bias. Also, any variables used to construct an interaction term must also exist in the model as a main effect to capture their individual contributions to the model. Table 8 shows that both of these conditions are met.

It is also important that the mixed model exhibit normally distributed residuals, since this distribution is assumed when using the maximum likelihood regression method (used by SAS in *Proc Mixed*). SAS automatically performs residual calculations and outputs several plots that may be used to assess their distribution. These plots are generated for the descriptive model shown in Figure 4. The residuals resulting from the descriptive model exhibit the desired bell-shape, but do not follow the expected distribution in the extremes. Rather, they exhibit a condition often referred to as "long



Residual Statistics

| | |
|---|---|
| Observations | 937 |
| Minimum | -7.703 |
| Mean | -76E-6 |
| Maximum | 5.5617 |
| Std Dev | 0.8751 |

Fit Statistics

| | |
|---|---|
| Objective | -8219 |
| AIC | -8191 |
| AICC | -8191 |
| BIC | -8160 |

**Figure 4. SAS Residual Plot Output for Descriptive Model**

tails." These long tails are due to several programs which are not predicted well by the model. As the model is improved, the majority of programs are more accurately predicted. But programs that were poorly-predicted initially are not improved, causing the residuals to become less normally distributed. This lack of normality makes model interpretation difficult, though it does not affect the model's power so long as the residuals are symmetrical. Also, as shown by the results, the error in these "poorly predicted" programs is below the level of practical significance, since their cost estimates are still greatly improved by the model. For example, Figure 4 shows that nearly all of the large residuals are for estimates from the SBIRS High program. After the model is used to correct program cost estimate error, SBIRS High still exhibits the largest single residual, but over 95 percent of the error in that estimate is eliminated.

## Descriptive Model CGF

In contrast to Figure 2, Figure 5 demonstrates that the descriptive model compensates for the vast majority of error in acquisition cost estimates for historical programs. The mean descriptive model-corrected CGF for all SARs is 1.0009, and the absolute model-corrected error is 0.4 percent. This represents, a 98.7 percent reduction of the error in the estimates. Figure 5 shows the estimate error plotted against *Percent Expended*; the axis is scaled to allow direct comparison with Figure 2. However, this coarse scale obscures any of the patterns in the model-corrected graph. The model-corrected error, plotted against *Percent Expended*, is shown again in Figure 6, with the axis constrained to ± 20 percent. This figure further illustrates the corrective power of the descriptive model. The only outliers on this graph are for the Space Based Infrared (SBIRS High) program. The rest of the observations are corrected to within 2 percent of the actual, final program cost.

Table 9 shows the summary for the descriptive model-corrected errors. The “Absolute Uncorrected” row lists the average of the absolute error with no model correction; the “Absolute Descriptive-Corrected” row shows the remaining absolute error after the descriptive model is used to correct the CGF.

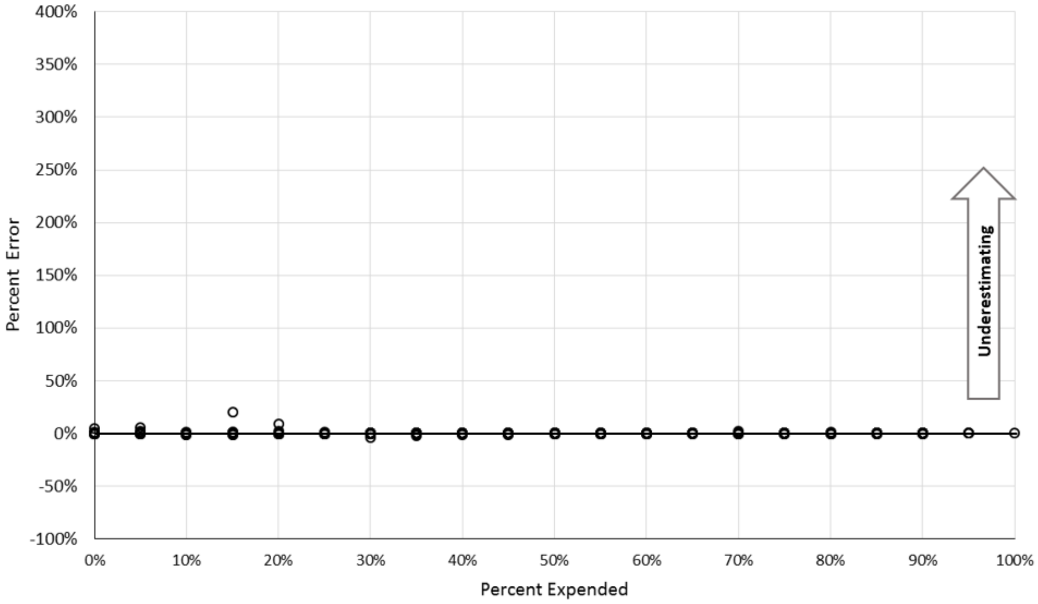


**Figure 5. Descriptive Model-Corrected Estimate Error by Percent Expended**

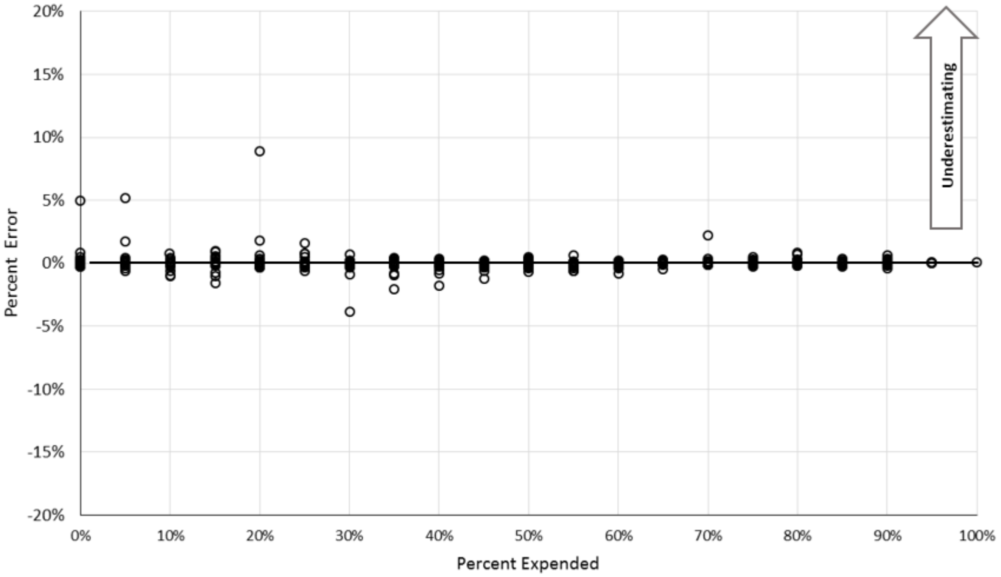
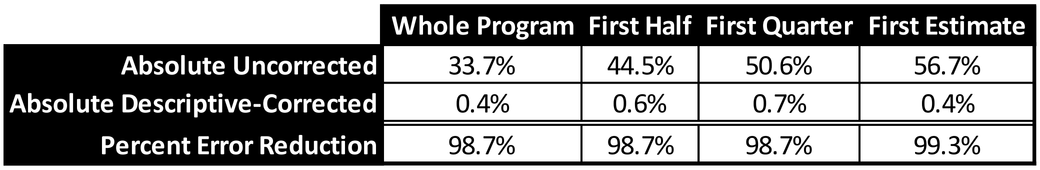


**Figure 6. Descriptive Model-Corrected Estimate Error by Percent Expended, Fine Axis Scale**

**Table 9. Average Descriptive Model Cost Estimate Error Reduction**



| | Whole Program | First Half | First Quarter | First Estimate |
|---|---|---|---|---|
| Absolute Uncorrected | 33.7% | 44.5% | 50.6% | 56.7% |
| Absolute Descriptive-Corrected | 0.4% | 0.6% | 0.7% | 0.4% |
| Percent Error Reduction | 98.7% | 98.7% | 98.7% | 99.3% |

Table 9 illustrates that the descriptive model reduces 98.7 percent of the error in cost estimates, on average, across program life (the “Whole Program” column). It also shows that an equivalent reduction is achieved if averaged over the first half, or first quarter of the estimates. In other words, the descriptive model is equally useful throughout the life of the program.

## Descriptive Model Confidence

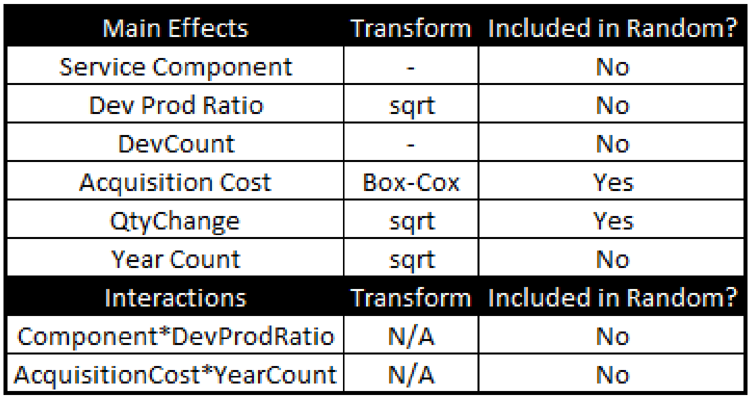
SAS has the capability to calculate the prediction intervals around the model-predicted value. However, these prediction intervals are formed around the transformed dependent variable and cannot be easily interpreted. When these intervals are transformed back into linear percentages, the non-linear transform causes the intervals to take values, in some cases, in excess of a thousand percent. Since this non-linear transform makes interpretation difficult, fixed ranges around the predicted value are examined to determine the percentage of true CGF values captured.

For the descriptive model, it is not surprising that the interval is very narrow. Bounds of just ±0.5 percent are enough to capture 91 percent of the true values. Bounds of ±1 percent capture 97 percent and bounds of ±2.5 percent capture 99 percent of the true values. These narrow bounds imply a high degree of confidence in the predictions from the descriptive model.

### Predictive Model

The results of the descriptive model illustrate the power of the macro-stochastic approach to cost estimating. However, this model is not useful for predicting future programs since the subject is the individual program, and the parameter estimates from a specific program cannot be extrapolated to a new program. Therefore, the regression is conducted once more using the bins that group similar programs according to their total cost growth score, as discussed in Chapter 3. Parameter combinations are tested using these Cost Growth Groups as the subject; the resulting model has six main effects and eleven terms, including the intercept. This model is summarized in Table 10.

**Table 10. Parameters in the Predictive Model**



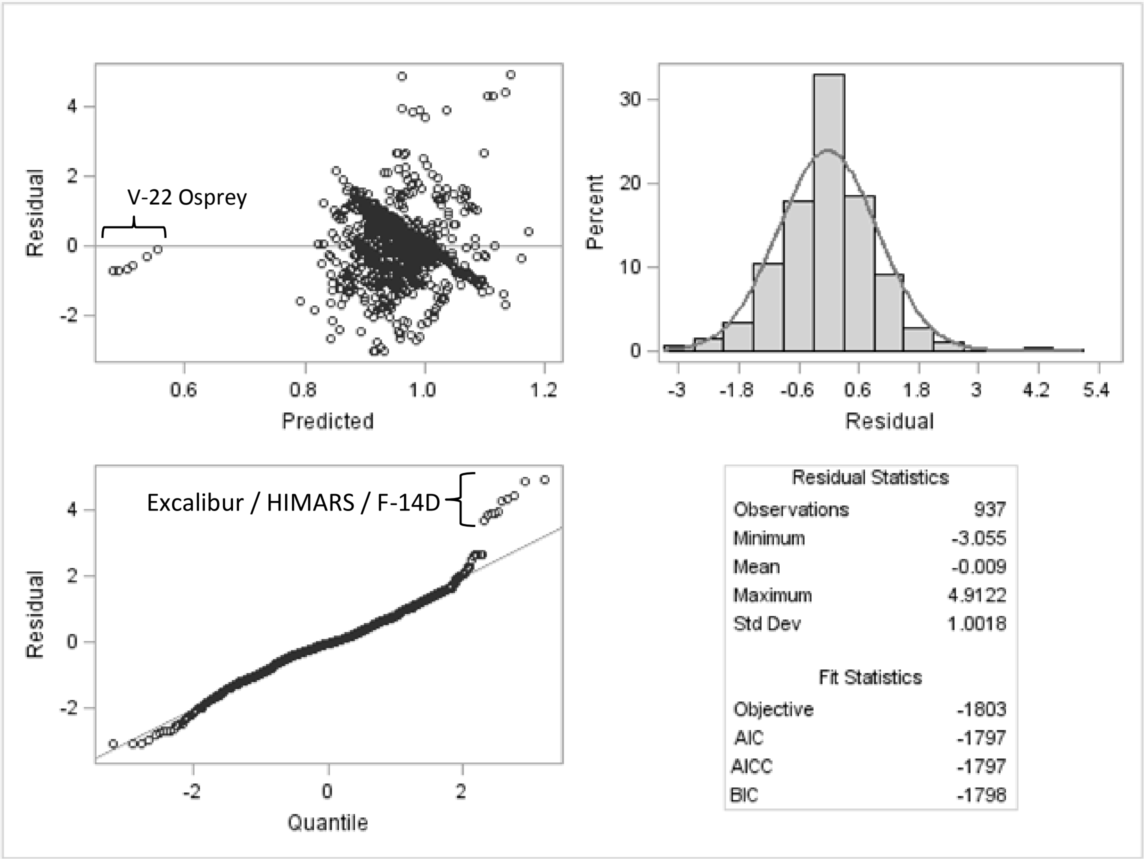
| Main Effects | Transform | Included in Random? |
|---|---|---|
| Service Component | - | No |
| Dev Prod Ratio | sqrt | No |
| DevCount | - | No |
| Acquisition Cost | Box-Cox | Yes |
| QtyChange | sqrt | Yes |
| Year Count | sqrt | No |
| Interactions | Transform | Included in Random? |
| Component*DevProdRatio | N/A | No |
| AcquisitionCost*YearCount | N/A | No |

The predictive model has two fewer main effects, and eight fewer terms than the descriptive model. There are several reasons for this difference. First, some of the parameters that are associated with the CGF in the descriptive model are used to bin programs into groups for the predictive model. For example, notice that Type, Iteration, and Years Funded are in the descriptive model, but not the predictive model. Since these terms make up three of the four variables that determine the program grouping, they still influence the inferences from model, but their parameter estimates are convolved with the intercept term and group-specific trajectories. Second, the descriptive model fits each program individually, which allows this model to accurately resolve the trends in each program. When several programs are combined into a group these trends may average out, obfuscating a once-meaningful relationship and replacing it with noise. This condition results in fewer meaningful parameters, and a greater chance of over-fitting the model. For these reasons, the size difference between the descriptive and predictive models is justified.

A first-order autoregressive—AR(1)—covariance structure best models the dependence within the data. Note that this is the same structure selected for the descriptive model. This structure results in a lower BIC for every examined combination of parameters. The parameter estimate value for the sequential correlation ($\rho$) is estimated and shown with the model outputs in Appendix A.

## Predictive Model Adequacy

As with the descriptive model, interactions are only composed of main effects, and all random effect terms are duplicates of a fixed effect. The SAS-generated residual plots are shown below in Figure 7.



Residual Statistics

| | |
|---|---|
| Observations | 937 |
| Minimum | -3.055 |
| Mean | -0.009 |
| Maximum | 4.9122 |
| Std Dev | 1.0018 |

Fit Statistics

| | |
|---|---|
| Objective | -1803 |
| AIC | -1797 |
| AICC | -1797 |
| BIC | -1798 |

**Figure 7. SAS Residual Plot Output for Predictive Model**

The residuals for the predictive model are normally distributed, with the exception of a cluster of high-residual observations. These points are a mixture of Excalibur (an Army munition program), HIMARS (an Army ground vehicle program), and F-14D (a

Naval aircraft program) observations, though not all observations for these programs are outlying. This mixture of qualitatively different programs seemingly prevents any categorical assessment from omitting these observations. The plot of residual versus predicted CGF points out another noteworthy data feature: the V-22 program is an extreme outlier with regards to uncorrected estimate error. However, this program is predicted very well, and most of the error in these estimates is eliminated, representing a 99.8 percent error reduction in the early program estimates.

### Predictive Model-Corrected Estimate Error

The predictive model predicts the CGF in each observation and this prediction is used to correct the original acquisition cost estimate, just as with the descriptive model. The results with this correction applied are shown in Figure 8. Notice that the predictive model-corrected estimates grow slightly better over time, as many of the parameters in the model (such as *DevCount*) are correlated with program duration. This figure should be compared directly with Figure 2, which shows the uncorrected estimate error plotted with the same axis dimensions.

Recall that the uncorrected mean absolute error was 34 percent. The predictive model has an overall mean error of 5.6 percent (underestimating), and the absolute error corrected by this model averages 20.4 percent, representing a 39.4 percent reduction from the average uncorrected cost estimate error. This model does not perform as well as the descriptive model since the individual cost error trajectory of each program is not modeled; rather, the trajectory of a group of programs is modeled.

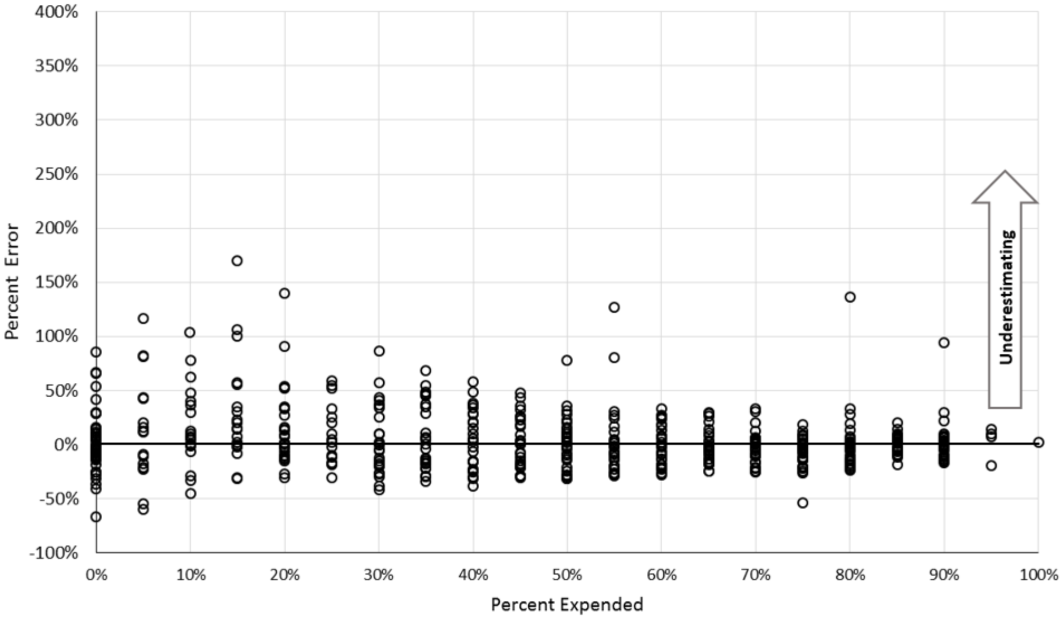
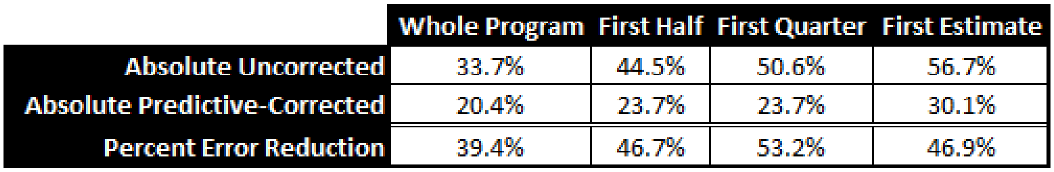


**Figure 8. Predictive Model-Corrected Estimate Error by Percent Expended**

Table 11 shows a summary of the predictive model performance. The performance of this model depends on the maturity of the estimate being corrected. For example, the predictive model reduces an average of 39.4 percent of the cost estimate error when applied to a cost estimate, selected at random. However, if this cost estimate is chosen at random from the first half of the program life, the error is reduced by 46.7 percent on average; if the estimate is chosen from the first quarter of program life, the error reduction is expected to be 53.2 percent, on average. This decreased utility over time is expected because, while the predictive model gets slightly better with repeated observations, the SAR estimate tends to converge to the actual cost as the program approaches completion.

**Table 11. Average Predictive Model Cost Estimate Error Reduction**



| | Whole Program | First Half | First Quarter | First Estimate |
|---|---|---|---|---|
| Absolute Uncorrected | 33.7% | 44.5% | 50.6% | 56.7% |
| Absolute Predictive-Corrected | 20.4% | 23.7% | 23.7% | 30.1% |
| Percent Error Reduction | 39.4% | 46.7% | 53.2% | 46.9% |

## Model Validation

The predictive model is capable of reducing estimate inaccuracy by more than half when used early in program life. However, these results are still not representative of a true prediction, since the program data for each of the predicted observations is used when fitting the model. The modified-LOOCV methodology, described in Chapter 3, is used to build 70 separate predictive models, each with a single program omitted. These models are then used to predict the correction factors for each SAR of each omitted program. The individual results from these 70 model validations are aggregated to estimate the predictive power of the model, and this aggregation is referred to as the "validated model" (though it technically represents 70 different validated models). The results for the validated model are shown in Figure 9, plotted with the same axis constraints as the previous figures in order to allow direct comparison.

The validated model has a mean of 12.6 percent (underestimating), and the average absolute error for the validated model is 27.4 percent. This error represents an 18.7 percent improvement from the uncorrected data, though the performance is more markedly improved when applied to earlier program estimates, as shown in Table 12.

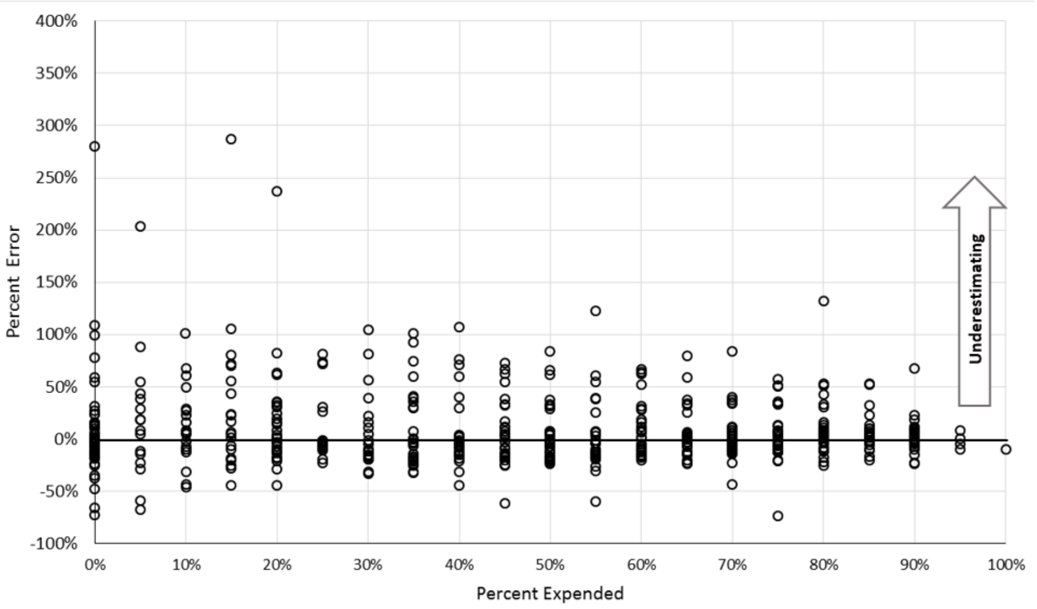
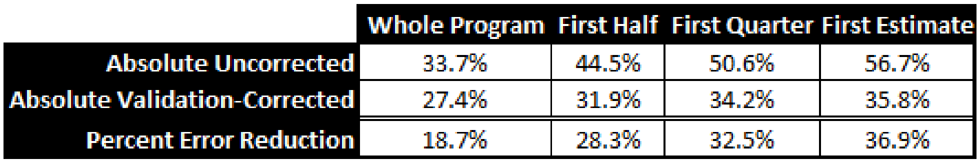


**Figure 9. Validation Model-Corrected Estimate Error by Percent Expended**

**Table 12. Average Validation Model Cost Estimate Error Reduction**



| | Whole Program | First Half | First Quarter | First Estimate |
|---|---|---|---|---|
| Absolute Uncorrected | 33.7% | 44.5% | 50.6% | 56.7% |
| Absolute Validation-Corrected | 27.4% | 31.9% | 34.2% | 35.8% |
| Percent Error Reduction | 18.7% | 28.3% | 32.5% | 36.9% |

## Validated Model Confidence

As with the validated model, the non-linear transformation on the dependent variable makes the customary prediction intervals difficult to interpret. The same fixed-bounds method is used to evaluate model confidence, though these intervals are expected to be wider than those for the descriptive model. Bounds of ±20 percent around the validation model CGF predictions capture 71 percent of the true CGF values. Bounds of ±35 percent capture 90 percent of the observed values and bounds of 45 percent capture 95 percent of the observed values. The usefulness of these bounds is discussed in Chapter 5.

### Validated Model Efficacy over Program Life

As explained above, the uncorrected estimate error trends towards zero as program acquisition nears completion, while the model predicting this error does not. For this reason, the predictive model is best used early in a program's life. Since the goal of this research is to provide a supplemental cost estimating tool for use early in program life, the efficacy of this tool is examined as a function of program life in Figure 10. Figure 10 shows the percentage of all estimates that are improved by the validation model, plotted against the percent of program expenditure, rounded to the nearest 5 percent. For example, the validation model improved 21 of the 29 SAR estimates (72 percent) produced when the program was approximately 5 percent expended. These data suggest a linear relationship, and this relationship—shown by the regression line in Figure 10—indicates that for each additional percent expended, the model loses nearly three-quarters of a percent of its predictive power. Equation 8, shown below, explains 88.4 percent of the variance—a strong relationship.

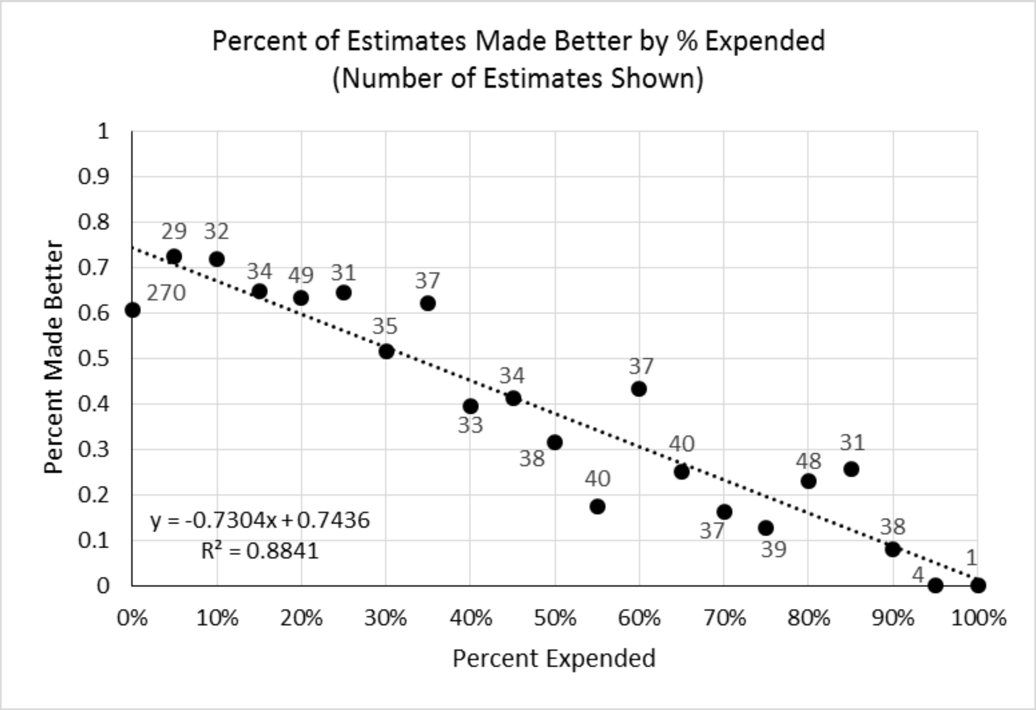


**Figure 10. Percent of Estimates Made Better with Validation Model, by Percent Expended**

$$\text{Percent of Estimates Improved} = 0.7436 - 0.7304 \cdot (\text{Percent Expended}) \quad (8)$$

This metric requires striking a balance between sample size for each average, and sample size for the linear regression. Figure 10 uses twenty-one data points to fit the demonstrated curve. These data points are created by rounding the program expenditure (reported to four decimal places) to the nearest five percent, but rounding up to the nearest ten percent, or down to the nearest 2.5 percent yields a similar equation and R-square. This tolerance indicates that the demonstrated relationship is robust, and not just an artifact of the rounding method.

## Model Efficacy by Program Cost Growth Group

Figure 11 shows the improvement in the absolute cost estimate errors, stratified by cost growth groups. Notice that the model performs much better against cost estimates with high amounts of initial error, as expected. Since the algorithm used to group programs incorporates parameters associated with cost growth, model efficacy can be improved by applying it only when it is expected to have a significant improvement on the estimate accuracy.

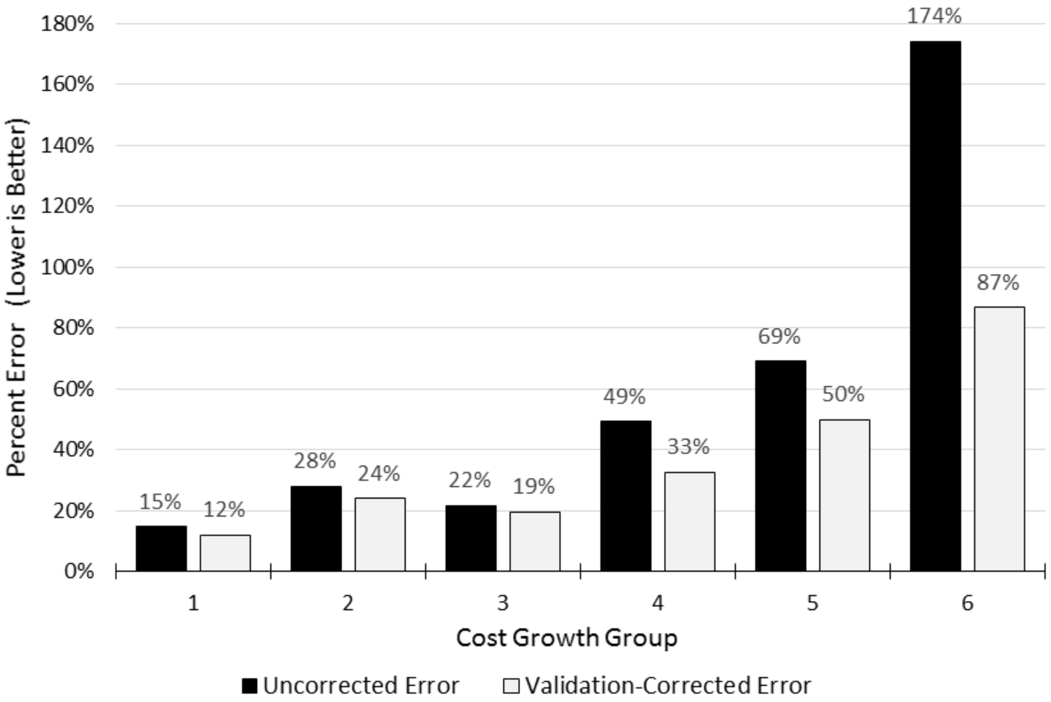


**Figure 11. Absolute Estimate Error by Cost Growth Group, Average over First Quarter of Program**

As shown in Figure 11, the difference in uncorrected and corrected error for the first three groups—collectively called *low-growth* programs—are all less than 5 percent. While this 5 percent represents a 15 percent reduction, on average, the relatively accurate initial estimates mean that the practical significance of the model is limited. In contrast, the difference for the latter three groups—collectively called *high-growth* programs—represent an error reduction of 38 percent on average, and as high as 50 percent in group 6. Model performance in these programs is more likely to be deemed practically significant by the user. Applying the model to the initial estimates for all 70 programs improves 49 of them (70 percent). If the model is applied only to the 30 initial estimates in cost growth groups four through six, 27 of them (90 percent) are improved.

## Considerations for Program Size

The metrics discussed above normalize the model results for program size by reporting estimate error as a percentage of the final acquisition cost. This normalization removes a meaningful result from model-corrected estimates. The goal of this research is to improve allocation of actual dollars and it is possible that model performance varies by program size. For instance, it's possible that the model improves small-dollar programs, but not high-dollar programs, resulting in overall poor performance in terms of absolute dollars. Using the Base Year 2013 estimated inflation rates, a metric is constructed which subtracts the model-corrected estimate error from the uncorrected estimate error, and converts this improvement to Base Year 2013 dollars. For example, consider an estimate for a $500 million program which is known to have 10 percent absolute error (an error of 50 million dollars). If the validated model corrects this error to only 5 percent (an error of 25 million dollars), then the model is said to have improved resource allocation by 25 million dollars. Note that the absolute error is what matters, since overestimation and underestimation are considered to damage resource availability by equal amounts.

If the validation model is applied to the first estimate of each program in the dataset, the total number of dollars reallocated is $91.0 billion (Base Year 2013). Due to the completion criteria placed on the data, there are no initial estimates after 2007, equating to an average of $4.3 Billion per year, when averaged over the 21 year span for the dataset. This reallocation does not mean that the DoD overspends by $4.3 Billion per year, but rather that this amount is inefficiently allocated due to the total effect of programs poorly estimating their actual resource needs by varying degrees.

Figure 12 shows that the model performs poorly for the smallest MDAPs—those with less than $2 billion in actual BY13 cost. These programs make up 13 of the 70 programs, about 19 percent. For MDAPs with a final cost between $2B and $5B, and those greater than $20B about 5 percent improvement is seen. These programs account for 27 of the 70 programs in the dataset, about 39 percent. The largest improvement is seen on the remaining 30 programs with between $5B and $20B in actual cost. Because the combined cost of the smallest programs is eclipsed by those in the larger categories (note the non-linear scale on the abscissa of Figure 13), the negative impact of these poorly predicted programs is minimized, as shown in Figure 13, resulting in a total improvement of 91 Billion BY13 dollars, reallocating approximately 9 percent of the $1.01 trillion dollar portfolio modeled by the dataset. If the sample of MDAPs in this

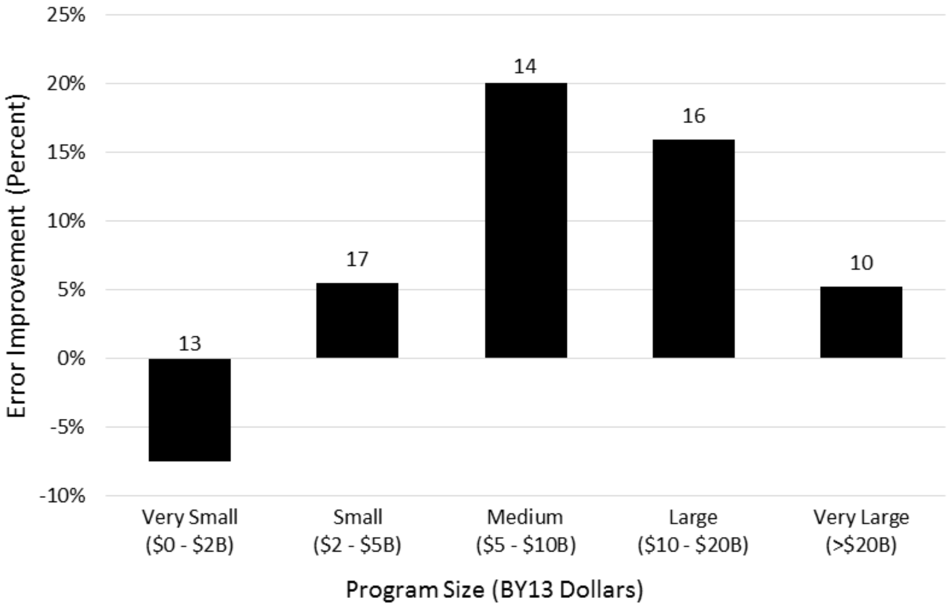


**Figure 12. Validation Model-Corrected Improvement, Percentage of Program Size**

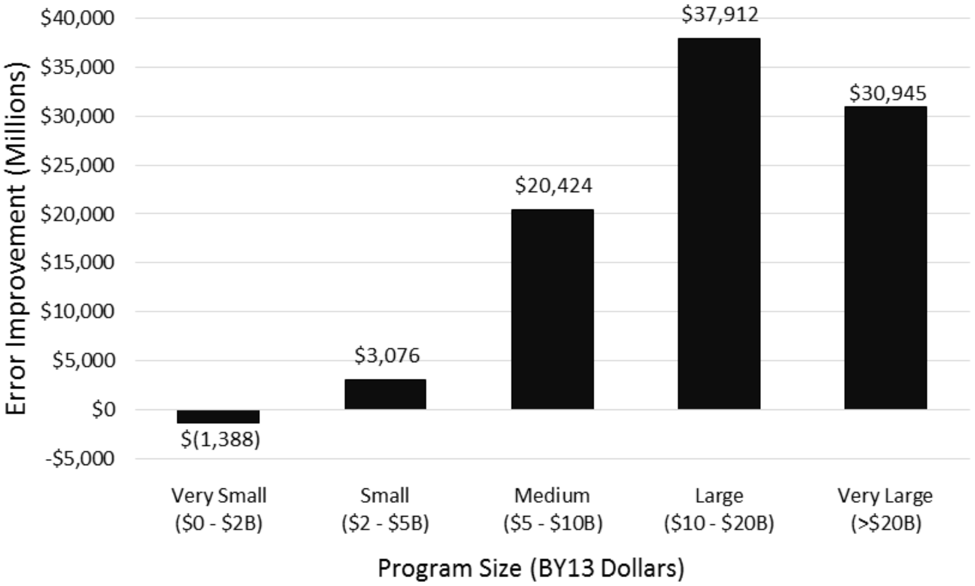


**Figure 13. Validation Model-Corrected Improvement in BY13 Dollars**

study is assumed similar to those in the current DoD MDAP portfolio (USD Comptroller, 2013), then this 9 percent reallocation equates to approximately $6.25 billion per year (BY13 dollars).

### Weighting Effects

The dataset includes 22 programs that are not complete, and these observations are weighted to reduce their influence on the regression parameters. The inclusion of a weighting methodology, explained in Chapter 3, lowers the BIC of the model, indicating that it is beneficial. However, when the model generated with weighted observations is compared to the same model generated with no weighting, the results are remarkably similar. The predictive model, generated without the weighting methodology, performs less than one percent worse, with 30.9 percent model-corrected error in the initial estimate. This difference in the average model performance with and without weighted observations is not practically significant.

The low impact of the weighting is likely due to the fact that incomplete programs make up less than a third of the total programs (22 of the 70) and the average maturity in these incomplete programs is 74 percent (measured by *Percent Expended*). Also, the observed absolute cost estimate error, measured in the final quarter of the program, is low, only about 8 percent. The error in the last third of program life—where expenditure is greater than 66 percent—is only about 10 percent. This low error implies that these incomplete but mature programs are already an adequate approximation of the final program cost, minimizing the impact of the weighting methodology.

## Chapter Summary

The validated predictive model is capable of significantly reducing the acquisition estimate error, even from the very first estimate. This meaningful result allows reallocation of 9 percent of the MDAP portfolio. Several trends are apparent that help to focus model usage, increasing its average performance even further. Of the 30 high-growth programs (programs in the upper three cost growth groups) 27 of these are improved by the validated model (90 percent). Also, the validated model performs well for all programs except the very smallest—those with a final expenditure of less than $2 billion (BY13). Finally, the model is best employed early in the program life, with the model losing one quarter percent of its efficacy for every additional percent expended. When the model is applied to the most favorable subset of the sample—the first estimate of a high-growth program with eventual cost over $2B—the average absolute error is reduced by approximately 45 percent.

The combined effect of using the validation model to correct all 70 initial cost estimates in the dataset is shown in Figure 14. The histogram of corrected and uncorrected CGFs, measured from the first estimate (such as the one in Figure 3, on page 56) are used to fit a Gamma distribution. The Kolmogorov-Smirnov test shows that both distributions in Figure 14 are acceptable matches at the $\alpha$= 0.05 level, despite the low sample size of only 70 data points used to fit these curves. The model-corrected distribution is shown to be more symmetrical, (implying a lower bias towards underestimation) with an average CGF closer to the desired value of 1.0, and a lower variance from this value. Using the model-corrected CGF, 38 of the 70 programs have a

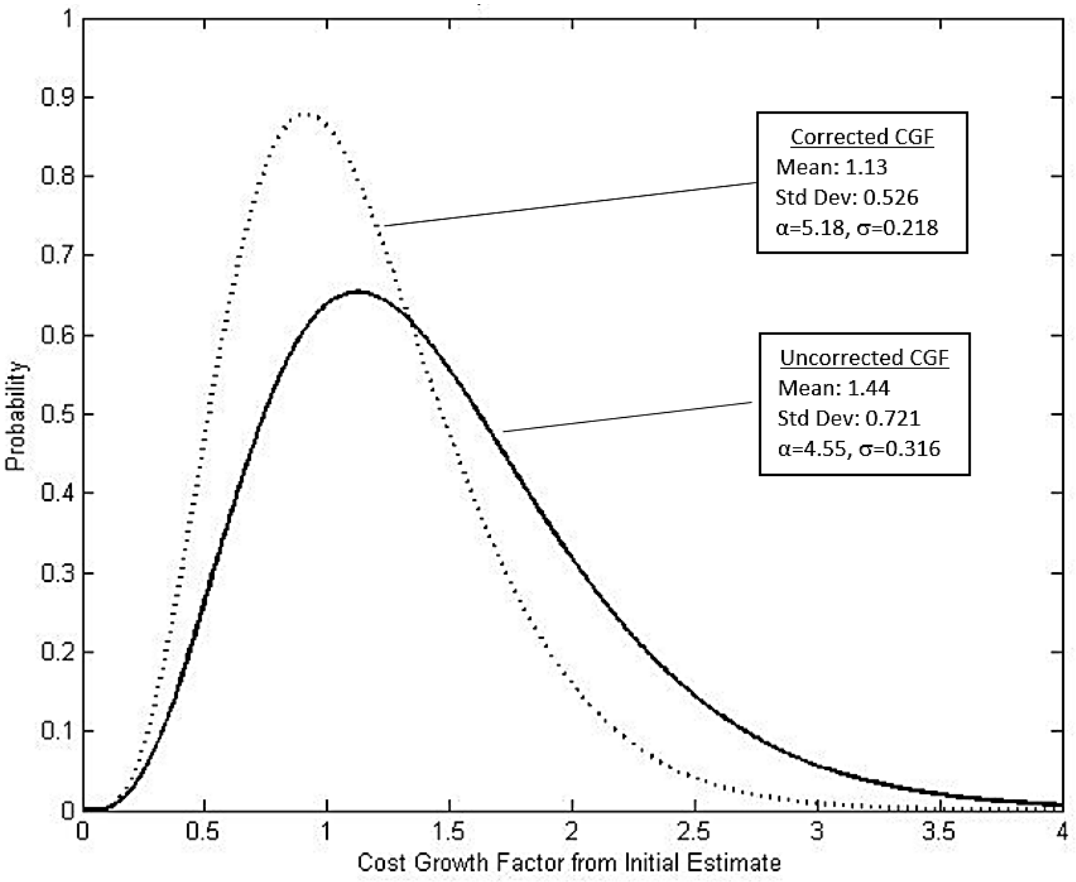


**Figure 14. Comparison of Fitted Gamma Distributions for Initial Estimate Error**

final cost within 25 percent of the estimate, compared to only 25 programs using the program office estimate. The difference in the distributions has a p-value of 0.011, significant at the α= 0.05 level.

# V. Conclusions and Recommendations

## Introduction

The macro-stochastic models are shown to reduce the errors in DoD acquisition cost estimates by meaningful amounts. However, it is necessary to discuss the assumptions behind these predictions, and highlight limitations for model use. This chapter revisits the research questions from Chapter 1, and answers them using the data from Chapter 4. Finally, questions resulting from this effort are presented in order to stimulate future work in the area of macro-stochastic cost estimation.

## Model Use

Chapter 4 summarizes model performance by reporting error reduction from different perspectives. For example, the model reduces more error on high-risk programs, and on programs with a final expenditure greater than $2 billion dollars (BY13). However, the average number of estimates improved by the model drops below 50 percent—the figurative “coin flip”—when programs are only 30 percent expended. This relationship also holds for the absolute average model-corrected error, which becomes equivalent to the program office error when a program is around one-third expended. This degradation in model performance implies a window for use of the validated prediction model. However, note that some of the variables in the predictive model (and therefore, in the validation model) measure some aspect of change in a program. For example, quantity change is a significant parameter, and this “change” variable cannot be measured in the first year. For this reason, the predictive model will sometimes improve over the first quarter of program life as shown in Table 11.

However, since estimate corrections suggested by the model after a program is approximately one-third expended are not expected to be significantly better than the original estimate, it should not be employed past this maturity threshold.

The confidence bounds placed around the predicted CGF show that 90 percent of the observations are captured within ±35 percent of the predicted value. This interval may seem wide, but 51 percent of initial estimates fall outside of this range. Of the 30 programs in the high-growth groups, 20 of them (66.6 percent) fall outside of the 35 percent confidence bounds. These intervals (which attempt to enclose the true value) should not be confused with the data presented in Figure 10 (which only attempts to make estimates better). In other words, even though the ±35 percent encloses the initial estimate about half of the time, 71 percent of initial estimates are made better by at least correcting in the direction indicated by the predicted CGF. Also, 90 percent of the programs in the high-growth program groups are made better by correcting in the direction indicated by the predicted CGF.

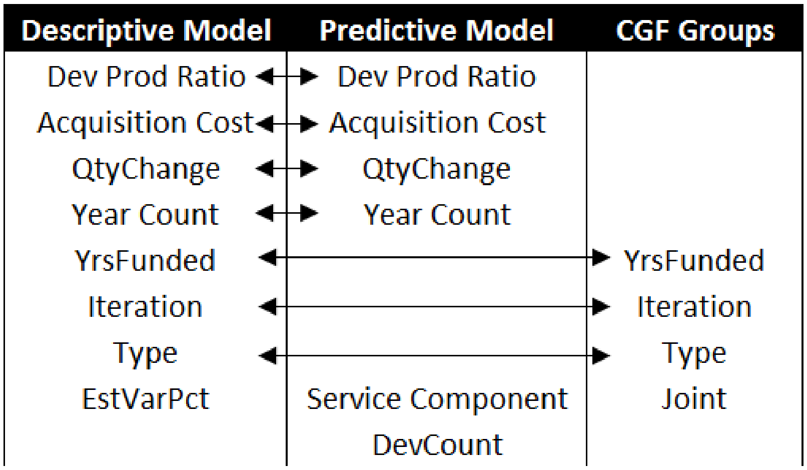
It is important to understand that while a few individual programs may be poorly predicted by the model due to erratic or unusual trends in their estimate errors, the purpose of the model is to inform resource allocation at the portfolio level where many programs will be monitored and *average* model performance is more relevant. As such, the models are not a justification for management reserve—a practice forbidden in DoD budgeting—nor is it a tool to assist program managers in identifying risks to their programs. In fact, since the underpinnings of the macro-stochastic approach rely upon correlation, not causation, intimate knowledge of the models by low-level decision makers could alter the nature of the observed relationships, rendering the models less effective or even useless.

Finally, the macro-stochastic model should not be used to drive acquisition reform or assign blame for cost growth. Since mixed models fit groups of programs to their unique estimate error trends, it can be misleading to evaluate the meaning of model parameters by examining the magnitude (and direction) of their coefficients. For example, the regression might show that program size (as measured by estimated acquisition cost) is uncorrelated with cost growth, but this relationship might be an average of some programs where the correlation is highly negative, and other programs where the exact opposite relationship holds true. Also, this acquisition cost variable would only be *correlated* with cost growth—that is, cost growth is almost certainly not *caused* by program size. Finally, transformations performed on some variables (such as the inverse cube root of acquisition cost) deter meaningful interpretation of the effect of these independent variables upon the dependent variable.

**Significant Parameters**

In addition to the interpretation difficulties imposed by random effects and variable transformations, the program grouping algorithm employed for the predictive model further obfuscates variable significance. Selection of significant predictors of cost growth to group programs into the six CGF groups convolves the effect of these parameters with the intercept term for the group. In other words, the parameters are present in the model, but it is not possible to determine their effects on the CGF individually. Interestingly, Table 13 shows that when these variables are considered alongside the list of predictive model parameters, the combined list is nearly a 90 percent match for the list of significant, descriptive model parameters. While we can say with confidence that these are among the most significant model parameters—recall that not all 3.77x$10^{22}$ combinations were tested—the significance level of a specific parameter cannot be determined.

**Table 13. Comparison of Model Parameters (Main Effects)**



| Descriptive Model | Predictive Model | CGF Groups |
|---|---|---|
| Dev Prod Ratio | Dev Prod Ratio | |
| Acquisition Cost | Acquisition Cost | |
| QtyChange | QtyChange | |
| Year Count | Year Count | |
| YrsFunded | | YrsFunded |
| Iteration | | Iteration |
| Type | | Type |
| EstVarPct | Service Component | Joint |
| | DevCount | |

## Impact of Key Assumptions

One of the key assumptions implicit in the selected first-order autoregressive AR(1) covariance structure is that of independence between programs. This assumption allows for the selection of simpler covariance structures, although known violations exist in this dataset. For example, Cancian's 2010 article mentions the Navy's cuts to the new DDG-1000 destroyer in favor of purchasing more of the older DDG-51 ships (Cancian, 2010)—both of these programs are in the dataset. Also, the Army's "Longbow" helicopter, and the “Longbow Hellfire” munition developed for that helicopter, are certainly correlated programs. In fact, mentioning these specific examples is perhaps myopic, given the likelihood that all programs are correlated to some degree since they are subject to the same economic, political and budgetary constraints. These *macro-*macro-level variables are beyond the scope of this study, though they likely play a role in driving cost estimate error.

Another assumption is that the program’s estimate of its final cost, generated at 90 percent completion, is acceptably accurate to allow correction to the unknown true cost. As explained in Chapter 2, Tracy’s 2005 study of “Estimate at Completion” indicates that estimates generated at 92.5 percent can be considered “final” costs (Tracy, 2005). However, it is not necessary to have the exact final cost in order to build a useful model. The programs in the database for this research exhibit approximately 10 percent absolute error in the last third of program life, and approximately 8 percent in the last quarter. The additional 2.5 percent expenditure over the required 90 percent would occur in less than a year in almost every program evaluated. Also, 22 of the 48 completed programs’ final estimates are generated with expenditures that exceed 90 percent, due to the annual report cycle. Therefore, the error in the 90 percent cost estimate is likely only a few percent different from the 92.5 percent estimate, except in a few rare cases where large changes were made at the very end of the program. The lack of effect from the program weights further bolsters this argument.

## Generalizability

One major limitation to generalizability is the similarity between the programs used in this study, and the rest of the DoD acquisition population. As discussed, only MDAPs that meet relatively restrictive filtering criteria are used in the analysis, and that fact reduces the generalizability of inferences drawn from the resulting model. This study represents a subset of all DoD acquisition programs and uses assumptions to overcome the myriad challenges in SAR analyses that are discussed in Chapter 2. Inferences drawn from this research should be limited to DoD programs that fall within the range of the filtering criterion. The results presented in Chapter 4 should not be used to draw inferences on non-MDAP programs, MAISs, or pre-Milestone B programs, as these may behave very differently. These filters notwithstanding, certain assumptions are made to expand the dataset and allow as many programs into the dataset as possible. For example, the program completion threshold is largely determined through logistical considerations (though it exceeds the completion thresholds used in other studies). Also, the weighting scheme, designed to reduce the influence of these incomplete programs, does not produce a meaningful difference on the parameter estimates, but expanding the dataset further may introduce programs that require weighting to reduce the effect on the model.

Extrapolation is also an issue. Table 1 illustrates that each nominal program variable used in the study has a sample size sufficient for robust estimation. However, the combination of multiple factors reduces this sample size. For example, inferences about Army programs may be drawn from a sample of 16 programs, while inferences about Army munition programs must be drawn from a sample of only three programs. Furthermore, all possible combinations of all factor-levels are not represented in the data. For example, using the model to predict the CGF of an Air Force ground vehicle might produce poor results, since no such programs exist in the dataset.

### Investigative Questions Answered

The three investigative questions from Chapter 1 may now be answered using the data from the analysis presented in Chapter 4.

1. **What program characteristics are the most significant predictors of acquisition cost growth?** As discussed above, it is difficult to attribute significance to predictors individually, but the list of identified predictors is similar between the predictive and descriptive models. Both models incorporate the program acquisition cost, the year count, the expected number of funding years, the program iteration, and the type of program. Additionally, the ratio of development to production years, and any changes in the procurement quantity are identified as significant in both models. The descriptive model includes the variance due to estimating differences, though this variable is not present in the predictive model. The predictive model includes the service component, joint status, and the number of development APBs, though none of these variables are present in the descriptive model.

2. **How can the selected factors be used to modulate the acquisition cost estimate, reducing the error?** As demonstrated in Chapter 4, correcting acquisition cost estimates is achieved by conducting a regression using the CGF predictors, and then multiplying this predicted CGF for some year by the estimate in that year. When applied to the first estimate, this methodology reduces cost estimate error by over a third. When applied in the most advantageous conditions, (applied early to high-risk, high-dollar programs)it reduces cost estimate error by nearly half.

3. **What level of confidence is achieved by predicting acquisition cost growth using significant factors that are available at program initiation?** Bounds with fixed half-widths are placed around the validated model-predicted CGFs to assess confidence. Bounds of ±35 percent capture 90 percent of the true values, and bounds of ±45 percent capture 95 percent of the true values. These ±35 percent bounds are sufficiently narrow that they do not enclose the initial program office estimate about half the time. While the other half of the initial estimates

are enclosed by these bounds, 71 percent of all estimates are improved by the validated model when using the predicted CGF to correct the estimate.

### Future Research

Many of the assumptions and limitations mentioned in this research may be used as inspiration for future research in the still-nascent area of macro-stochastic estimation. Several of these future research ideas are presented below.

The completion criteria placed on programs in the dataset do not require programs to be complete, since this would reduce the sample size by a third. However, a similar analysis performed on the program development phase would consider programs to be complete when they reach Milestone C, vastly increasing the number of programs eligible for analysis, and allowing more recent programs into the study. Furthermore, increasing the scope of the study to include economic and political indicators could increase estimate accuracy and predictive validity of the model.

This research uses the “mixed” procedure available in SAS 9.3 to perform the mixed-model regression. However, the more flexible generalized linear mixed model procedure known as the *glimmix* procedure is also available that uses different methods to optimize the parameter estimates. Preliminary examination of the dataset with this tool shows that it produces slightly different results, but allows some of the more complex covariance structures to converge. The “unstructured” covariance structure, for example, could possibly account for some of the correlations between programs, reducing the variance and increasing model power. However this structure frequently would not converge when using the mixed procedure. Using *glimmix* would likely be necessary for any analyses that attempts to analyze larger datasets with fewer simplifying assumptions.

Finally, the assumption of program “completion” might be analyzed by extracting cost at completion from DCARC to refine final estimates.  As stated above, these results are unlikely to produce considerably different results, but it could increase the face-validity of the method to stakeholders and decision makers, while also serving as a validation for recent work on estimate accuracy near completion.

## Conclusion

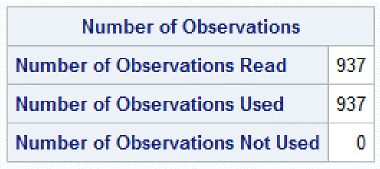
The macro-stochastic technique provides an economically and statistically meaningful improvement over initial program estimates, reducing cost errors by 18.7 percent, on average, when applied to any of the estimates for the programs in this dataset. However, the most logical usage for the model is to apply it to the initial estimate, and then utilize it to assist affordability decisions over the first third of program life when traditional estimates of program cost are at their worst.  If future programs are expected to perform similarly to those from the recent past, then this initial application of the model is expected to guide a more efficient allocation of about 9 percent of the MDAP portfolio—approximately $6.24 billion, annually.  Such a tool could prove invaluable to high-level decision makers and acquisition authorities who must make assessments of programs’ affordability based on little knowledge of its true cost.  In the current environment of budgetary reduction, efficient allocation of acquisition resources is crucial—macro-stochastic cost estimation is an excellent tool for this application.

# Appendix A

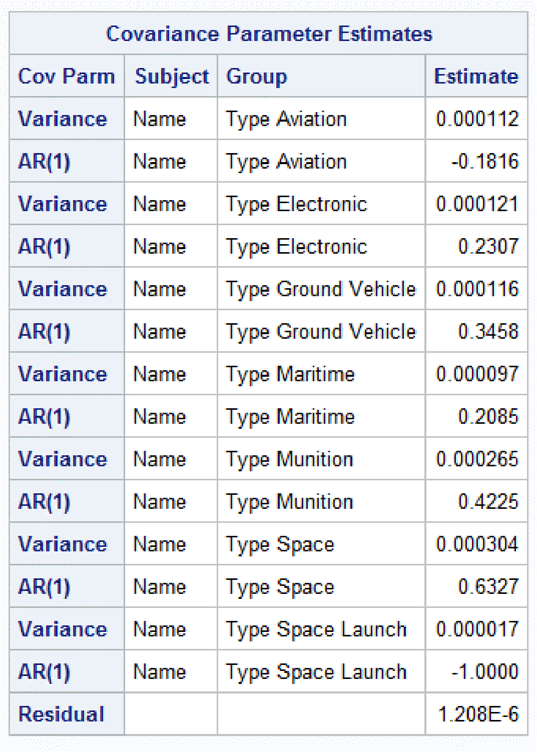
## Descriptive Model SAS Code

```
proc import out=work.mdaps
datafile="D:\Documents\School\THESIS\MDAPsNew.xlsx"
DBMS=XLSX;
Run;

data mdaps2; set mdaps;
logAcqCost=log(AcqCost);
sqrtAcqCost=sqrt(AcqCost);
sqrtProdCount=sqrt(ProdCount);
logFEE=log(FEE);
logQtyChange=log(QtyChange);
logYearCount=log(YearCount);
sqrtDPR=sqrt(Dev_Prod_Ratio);
logYrsFunded_i=log(YrsFunded_i);
BoxFee=Fee**(-1/3);
run;


ods html newfile=proc;
ods graphics on;
proc mixed data=mdaps2;

class Name iter type;

Weight Weight;

model BoxFEE=iter type iter*type sqrtDPR sqrtAcqCost EstVarPct
      logQTYChange logYearCount type*logQTYChange YrsFunded_i/solution
      residual OUTP=mdaps2Output;

Random int iter type QTYChange type*logQTYchange EstVarPct sqrtDPR
      sqrtAcqCost/sub=Name group=Type type=AR(1);
run;
Quit;
ods graphics off;

proc export data=work.mdaps2output
OUTFILE="D:\Documents\School\THESIS\MDAPsOutDesc_FINAL"
dbms=csv replace;
Run;
```

## Descriptive Model Outputs

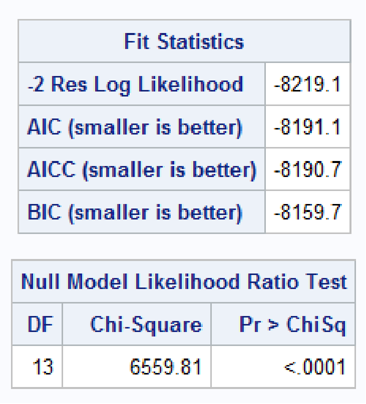


| Number of Observations | |
|---|---|
| Number of Observations Read | 937 |
| Number of Observations Used | 937 |
| Number of Observations Not Used | 0 |



Covariance Parameter Estimates

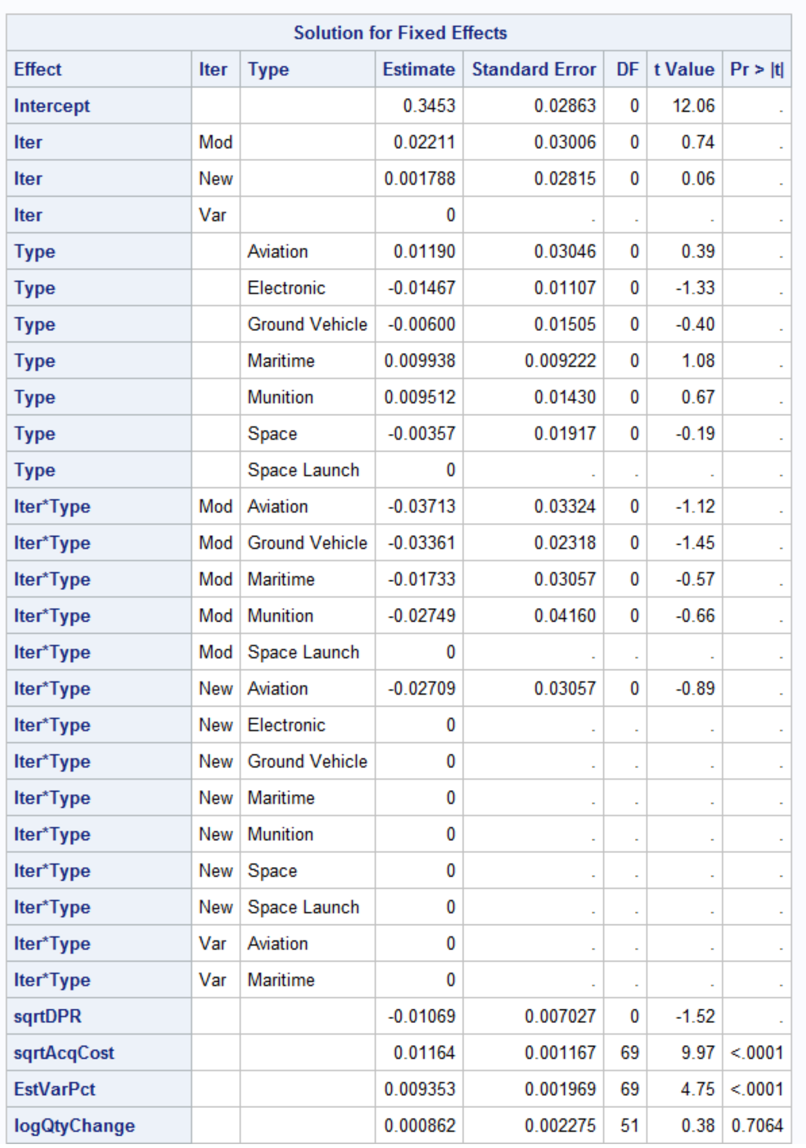
| Cov Parm | Subject | Group | Estimate |
|---|---|---|---|
| Variance | Name | Type Aviation | 0.000112 |
| AR(1) | Name | Type Aviation | -0.1816 |
| Variance | Name | Type Electronic | 0.000121 |
| AR(1) | Name | Type Electronic | 0.2307 |
| Variance | Name | Type Ground Vehicle | 0.000116 |
| AR(1) | Name | Type Ground Vehicle | 0.3458 |
| Variance | Name | Type Maritime | 0.000097 |
| AR(1) | Name | Type Maritime | 0.2085 |
| Variance | Name | Type Munition | 0.000265 |
| AR(1) | Name | Type Munition | 0.4225 |
| Variance | Name | Type Space | 0.000304 |
| AR(1) | Name | Type Space | 0.6327 |
| Variance | Name | Type Space Launch | 0.000017 |
| AR(1) | Name | Type Space Launch | -1.0000 |
| Residual | | | 1.208E-6 |



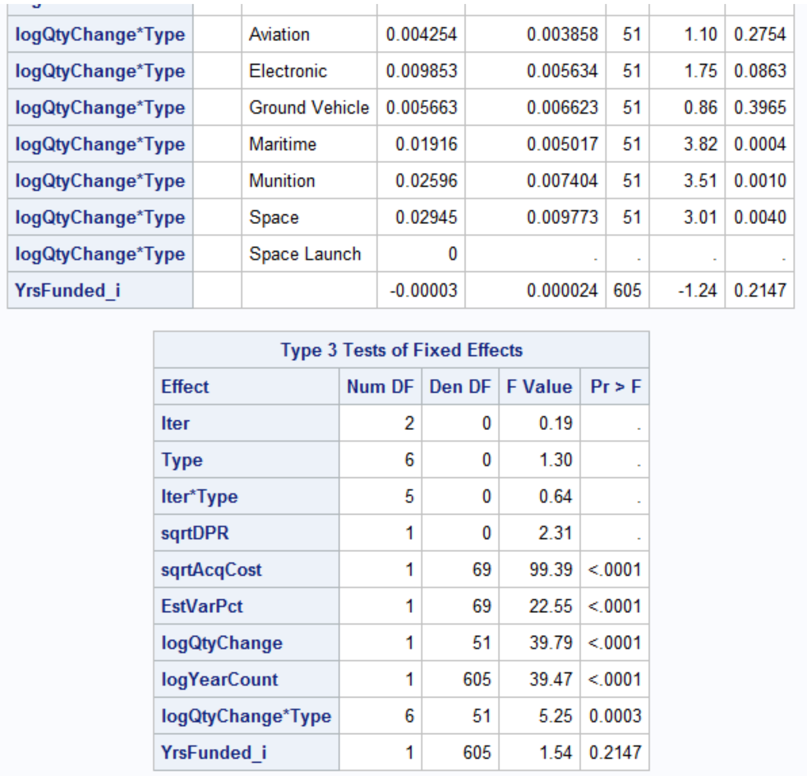
| Fit Statistics | |
|---|---|
| -2 Res Log Likelihood | -8219.1 |
| AIC (smaller is better) | -8191.1 |
| AICC (smaller is better) | -8190.7 |
| BIC (smaller is better) | -8159.7 |

Null Model Likelihood Ratio Test

| DF | Chi-Square | Pr > ChiSq |
|---|---|---|
| 13 | 6559.81 | <.0001 |



| Solution for Fixed Effects | | | | | | | |
|---|---|---|---|---|---|---|---|
| Effect | Iter | Type | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
| Intercept | | | 0.3453 | 0.02863 | 0 | 12.06 | . |
| Iter | Mod | | 0.02211 | 0.03006 | 0 | 0.74 | . |
| Iter | New | | 0.001788 | 0.02815 | 0 | 0.06 | . |
| Iter | Var | | 0 | . | . | . | . |
| Type | | Aviation | 0.01190 | 0.03046 | 0 | 0.39 | . |
| Type | | Electronic | -0.01467 | 0.01107 | 0 | -1.33 | . |
| Type | | Ground Vehicle | -0.00600 | 0.01505 | 0 | -0.40 | . |
| Type | | Maritime | 0.009938 | 0.009222 | 0 | 1.08 | . |
| Type | | Munition | 0.009512 | 0.01430 | 0 | 0.67 | . |
| Type | | Space | -0.00357 | 0.01917 | 0 | -0.19 | . |
| Type | | Space Launch | 0 | . | . | . | . |
| Iter*Type | Mod | Aviation | -0.03713 | 0.03324 | 0 | -1.12 | . |
| Iter*Type | Mod | Ground Vehicle | -0.03361 | 0.02318 | 0 | -1.45 | . |
| Iter*Type | Mod | Maritime | -0.01733 | 0.03057 | 0 | -0.57 | . |
| Iter*Type | Mod | Munition | -0.02749 | 0.04160 | 0 | -0.66 | . |
| Iter*Type | Mod | Space Launch | 0 | . | . | . | . |
| Iter*Type | New | Aviation | -0.02709 | 0.03057 | 0 | -0.89 | . |
| Iter*Type | New | Electronic | 0 | . | . | . | . |
| Iter*Type | New | Ground Vehicle | 0 | . | . | . | . |
| Iter*Type | New | Maritime | 0 | . | . | . | . |
| Iter*Type | New | Munition | 0 | . | . | . | . |
| Iter*Type | New | Space | 0 | . | . | . | . |
| Iter*Type | New | Space Launch | 0 | . | . | . | . |
| Iter*Type | Var | Aviation | 0 | . | . | . | . |
| Iter*Type | Var | Maritime | 0 | . | . | . | . |
| sqrtDPR | | | -0.01069 | 0.007027 | 0 | -1.52 | . |
| sqrtAcqCost | | | 0.01164 | 0.001167 | 69 | 9.97 | <.0001 |
| EstVarPct | | | 0.009353 | 0.001969 | 69 | 4.75 | <.0001 |
| logQtyChange | | | 0.000862 | 0.002275 | 51 | 0.38 | 0.7064 |



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| logQtyChange*Type | | Aviation | 0.004254 | 0.003858 | 51 | 1.10 | 0.2754 |
| logQtyChange*Type | | Electronic | 0.009853 | 0.005634 | 51 | 1.75 | 0.0863 |
| logQtyChange*Type | | Ground Vehicle | 0.005663 | 0.006623 | 51 | 0.86 | 0.3965 |
| logQtyChange*Type | | Maritime | 0.01916 | 0.005017 | 51 | 3.82 | 0.0004 |
| logQtyChange*Type | | Munition | 0.02596 | 0.007404 | 51 | 3.51 | 0.0010 |
| logQtyChange*Type | | Space | 0.02945 | 0.009773 | 51 | 3.01 | 0.0040 |
| logQtyChange*Type | | Space Launch | 0 | . | . | . | . |
| YrsFunded_i | | | -0.00003 | 0.000024 | 605 | -1.24 | 0.2147 |

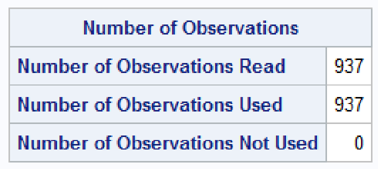
Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|
| Iter | 2 | 0 | 0.19 | . |
| Type | 6 | 0 | 1.30 | . |
| Iter*Type | 5 | 0 | 0.64 | . |
| sqrtDPR | 1 | 0 | 2.31 | . |
| sqrtAcqCost | 1 | 69 | 99.39 | <.0001 |
| EstVarPct | 1 | 69 | 22.55 | <.0001 |
| logQtyChange | 1 | 51 | 39.79 | <.0001 |
| logYearCount | 1 | 605 | 39.47 | <.0001 |
| logQtyChange*Type | 6 | 51 | 5.25 | 0.0003 |
| YrsFunded_i | 1 | 605 | 1.54 | 0.2147 |

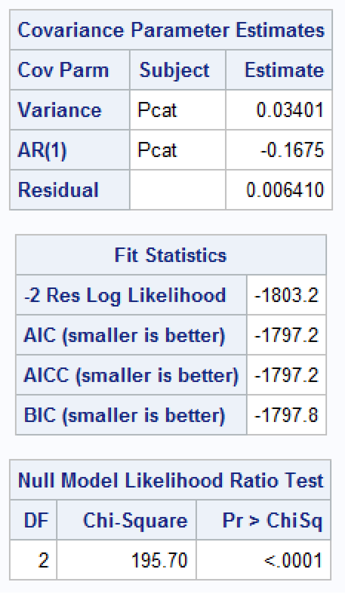
## Predictive Model SAS Code

```
proc import out=work.mdaps
datafile="D:\Documents\School\THESIS\MDAPsNew.xlsx"
DBMS=XLSX;
Run;

data mdaps2; set mdaps;
logQtyChange=log(QtyChange);
sqrtQtyChange=sqrt(QtyChange);
logFEE=log(FEE);
logYearCount=log(YearCount);
sqrtYearCount=sqrt(YearCount);
logYrsFunded_i=log(YrsFunded_i);
sqrtYrsFunded_i=sqrt(YrsFunded_i);
sqrtDevCount=sqrt(devcount);
sqrtProdCount=sqrt(prodcount);
sqrtYearCount=sqrt(YearCount);
sqrtAcqCost=sqrt(AcqCost);
logAcqCost=log(AcqCost);
sqrtUCBreachCum=sqrt(UCBreachCum);
sqrtDPR=sqrt(Dev_Prod_Ratio);
BoxFee=Fee**(-1/3);
BoxAcqCost=AcqCost**(-1/2);
run;
quit;

ods tagsets.excelxp file='Pred_allobs.xls' STYLE=statistical
            options( embedded_titles='yes' sheet_interval='proc' );
ods html newfile=proc;
ods graphics on;
title "Full Predictive Model";

proc mixed data=mdaps2;
class comp PCat;

model BoxFEE = comp sqrtDPR comp*sqrtDPR devcount BoxAcqCost
      sqrtQtyChange sqrtYearCount BoxAcqCost*sqrtYearCount/solution
      residual OUTP=mdaps2Output;

Random int logAcqCost sqrtQtyChange/sub=PCat type=AR(1) Solution;

Weight Weight;

run;
      ods tagsets.excelxp close;
quit;
ods graphics off;

proc export data=work.mdaps2output
OUTFILE="D:\Documents\School\THESIS\PredFINAL"
dbms=csv replace;
Run;
```

## Predictive Model Outputs



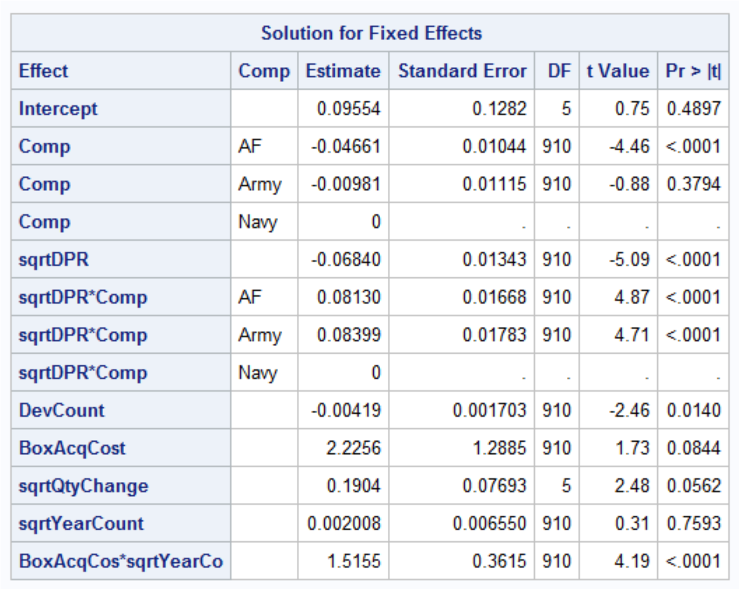
| Number of Observations | |
|---|---|
| Number of Observations Read | 937 |
| Number of Observations Used | 937 |
| Number of Observations Not Used | 0 |



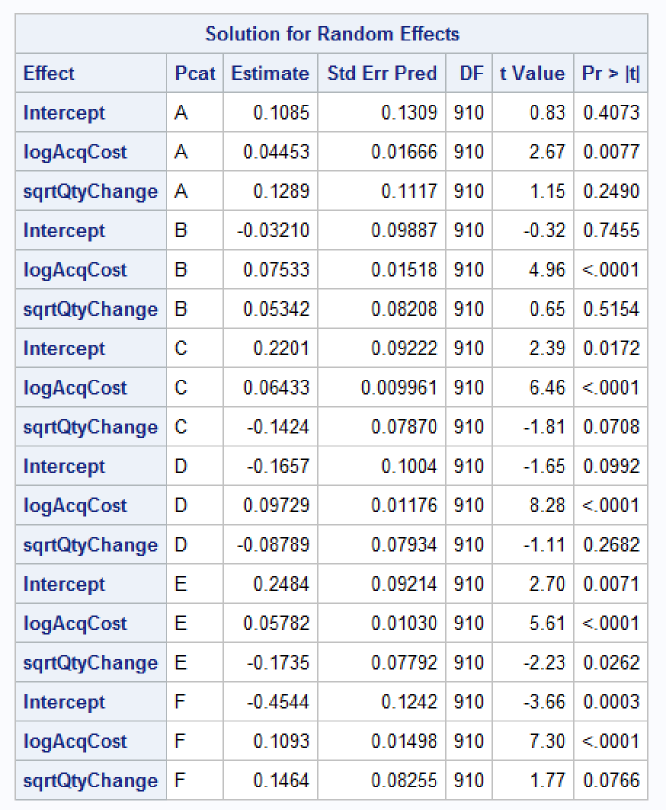
| Covariance Parameter Estimates | | |
|---|---|---|
| Cov Parm | Subject | Estimate |
| Variance | Pcat | 0.03401 |
| AR(1) | Pcat | -0.1675 |
| Residual | | 0.006410 |

| Fit Statistics | |
|---|---|
| -2 Res Log Likelihood | -1803.2 |
| AIC (smaller is better) | -1797.2 |
| AICC (smaller is better) | -1797.2 |
| BIC (smaller is better) | -1797.8 |

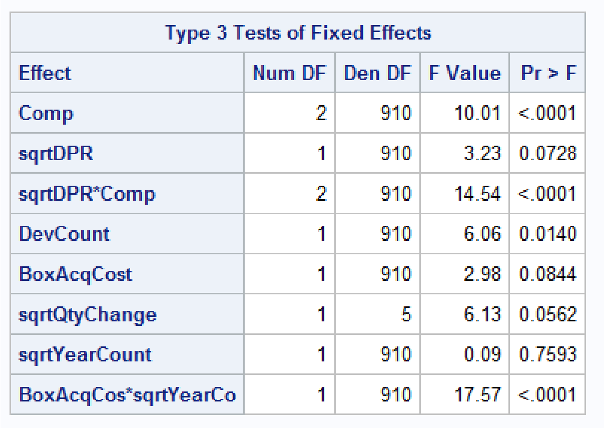
| Null Model Likelihood Ratio Test | | |
|---|---|---|
| DF | Chi-Square | Pr > ChiSq |
| 2 | 195.70 | <.0001 |



| Solution for Fixed Effects | | | | | | |
|---|---|---|---|---|---|---|
| Effect | Comp | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
| Intercept | | 0.09554 | 0.1282 | 5 | 0.75 | 0.4897 |
| Comp | AF | -0.04661 | 0.01044 | 910 | -4.46 | <.0001 |
| Comp | Army | -0.00981 | 0.01115 | 910 | -0.88 | 0.3794 |
| Comp | Navy | 0 | . | . | . | . |
| sqrtDPR | | -0.06840 | 0.01343 | 910 | -5.09 | <.0001 |
| sqrtDPR*Comp | AF | 0.08130 | 0.01668 | 910 | 4.87 | <.0001 |
| sqrtDPR*Comp | Army | 0.08399 | 0.01783 | 910 | 4.71 | <.0001 |
| sqrtDPR*Comp | Navy | 0 | . | . | . | . |
| DevCount | | -0.00419 | 0.001703 | 910 | -2.46 | 0.0140 |
| BoxAcqCost | | 2.2256 | 1.2885 | 910 | 1.73 | 0.0844 |
| sqrtQtyChange | | 0.1904 | 0.07693 | 5 | 2.48 | 0.0562 |
| sqrtYearCount | | 0.002008 | 0.006550 | 910 | 0.31 | 0.7593 |
| BoxAcqCos*sqrtYearCo | | 1.5155 | 0.3615 | 910 | 4.19 | <.0001 |



| Solution for Random Effects | | | | | | |
|---|---|---|---|---|---|---|
| Effect | Pcat | Estimate | Std Err Pred | DF | t Value | Pr > \|t\| |
| Intercept | A | 0.1085 | 0.1309 | 910 | 0.83 | 0.4073 |
| logAcqCost | A | 0.04453 | 0.01666 | 910 | 2.67 | 0.0077 |
| sqrtQtyChange | A | 0.1289 | 0.1117 | 910 | 1.15 | 0.2490 |
| Intercept | B | -0.03210 | 0.09887 | 910 | -0.32 | 0.7455 |
| logAcqCost | B | 0.07533 | 0.01518 | 910 | 4.96 | <.0001 |
| sqrtQtyChange | B | 0.05342 | 0.08208 | 910 | 0.65 | 0.5154 |
| Intercept | C | 0.2201 | 0.09222 | 910 | 2.39 | 0.0172 |
| logAcqCost | C | 0.06433 | 0.009961 | 910 | 6.46 | <.0001 |
| sqrtQtyChange | C | -0.1424 | 0.07870 | 910 | -1.81 | 0.0708 |
| Intercept | D | -0.1657 | 0.1004 | 910 | -1.65 | 0.0992 |
| logAcqCost | D | 0.09729 | 0.01176 | 910 | 8.28 | <.0001 |
| sqrtQtyChange | D | -0.08789 | 0.07934 | 910 | -1.11 | 0.2682 |
| Intercept | E | 0.2484 | 0.09214 | 910 | 2.70 | 0.0071 |
| logAcqCost | E | 0.05782 | 0.01030 | 910 | 5.61 | <.0001 |
| sqrtQtyChange | E | -0.1735 | 0.07792 | 910 | -2.23 | 0.0262 |
| Intercept | F | -0.4544 | 0.1242 | 910 | -3.66 | 0.0003 |
| logAcqCost | F | 0.1093 | 0.01498 | 910 | 7.30 | <.0001 |
| sqrtQtyChange | F | 0.1464 | 0.08255 | 910 | 1.77 | 0.0766 |



| Type 3 Tests of Fixed Effects | | | | |
|---|---|---|---|---|
| Effect | Num DF | Den DF | F Value | Pr > F |
| Comp | 2 | 910 | 10.01 | <.0001 |
| sqrtDPR | 1 | 910 | 3.23 | 0.0728 |
| sqrtDPR*Comp | 2 | 910 | 14.54 | <.0001 |
| DevCount | 1 | 910 | 6.06 | 0.0140 |
| BoxAcqCost | 1 | 910 | 2.98 | 0.0844 |
| sqrtQtyChange | 1 | 5 | 6.13 | 0.0562 |
| sqrtYearCount | 1 | 910 | 0.09 | 0.7593 |
| BoxAcqCos*sqrtYearCo | 1 | 910 | 17.57 | <.0001 |

### Validation SAS Code

```
ODS graphics on;
%MACRO sqlloop;
PROC SQL;
      Create Table prognames as
      Select Distinct Name from mdaps2
            ORDER BY Name;
      select count(*) into :nobs from prognames;
Quit;

ods tagsets.excelxp file='Validation.xls' STYLE=statistical
            options( embedded_titles='yes' sheet_interval='proc' );
%DO i=1 %TO &nobs;
*This takes each name and places it into a variable;
      PROC SQL noprint;
      select Name into :names from prognames(firstobs=&i obs=&i);
      Quit;

*this deletes the program with the currently selected name;
      data validate; set mdaps2;
      IF Name="&names" THEN Delete;
      Run;

*Here is the regression code from below, running without the selected 
program;
      title "&names";
      proc mixed data=validate noitprint noclprint noinfo;
      class comp PCat;
      model BoxFEE = comp sqrtDPR comp*sqrtDPR devcount BoxAcqCost 
            sqrtQtyChange sqrtYearCount BoxAcqCost*sqrtYearCount 
            /solution residual OUTP=mdaps2Output;
      Random int logAcqCost sqrtQtyChange/sub=PCat type=AR(1) Solution;
      Weight Weight;
      run;
*Now we output to a workbook, with a sheet named after the omitted 
program;
      %END;
      ods tagsets.excelxp close;
%MEND;


dm log 'clear' output;
%sqlloop;
```

## Bibliography

Arena, M. V., Leonard, R. S., Murray, S. E., & Younossi, O. (2006). *Historical Cost Growth of Completed Weapon System Programs.* Santa Monica, CA: RAND Corporation.

Box, G. E., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society, B*(26), 211-252.

Cancian, M. F. (2010). Cost Growth: Perception and Reality. *Defense Acquisition Research Journal*.

Defense Acquisition Management Information Retrieval System. (2014, January 13). http://www.acq.osd.mil/damir/

Defense Acquisition University. (2013). *Defense Acquisition Guidebook*. Retrieved January 15, 2014, from https://dag.dau.mil/Pages/Default.aspx

Drezner, J. A., Jarvaise, J. M., Hess, R. W., Hough, P. G., & Norton, D. (1993). *An Analysis of Weapon System Cost Growth* (MR-291-AF). Santa Monica, CA: RAND Corporation

Ferguson, D. R., et al. (2011). *Quantifying Uncertainty in Early Lifecycle Cost Estimation* (CMU/SEI-TR-025). Carnegie Mellon University.

Government Accountability Office. (2008). *A Knowledge-Based Funding Approach Could Improve Major Weapon System Program Outcomes* (GAO-08-619). Washington: GAO.

Government Accountability Office. (2009). *Cost Estimating Guide.* Washington: GAO.

Government Accountability Office. (2012a). *Defense Acquisitions: Assessments of Selected Weapon Programs* (GAO-12-400SP). Washington: GAO.

Government Accountability Office. (2012b). *Improvements Needed to Enhance Oversight of Estimated Long-Term Costs for Operating and Supporting Major Weapon Systems* (GAO-12-340). Washington: GAO.

Hough, P. G. (1992). *Pittfalls in Calculating Cost Growth from Selected Acquisition Reports* (N-3136-AF). Santa Monica, CA: RAND Corporation.

Jarvaise, J. M., Drezner, J. A., & Norton, D. (1996). *The Defense System Cost Performance Database.* Santa Monica, CA: RAND Corporation.

Kadish, R., et al. (2005). *Defense Acquisition Performance Assessment.* Washington: GPO.

Kincaid, C. (2005). Guidelines for Selecting the Covariance Structure in Mixed Model Analysis. *30th SAS User Group International.* Philadelphia, PA.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied Linear Regression Models.* New York: McGraw-Hill.

Office of the Undersecretary of Defense, Comptroller. (2013). *National Defense Budget Estimates for FY 2014.* Washington: GPO.

Patetta, M. (2002). *Longitudinal Data Analysis with Discrete and Continuous Responses Course Notes.* Cary, NC: SAS Institute.

Ryan, E. (2012, September). *Cost-Based Decision Model for Valuing System Design Options, Dissertation.* Wright-Patterson AFB, OH: Air Force Institute of Technology.

Ryan, E., Schubert-Kabban, C., Jacques, D., & Ritschel, J. (2013). A Macro-Stochastic Model for Improving the Accuracy of Department of Defense Life Cycle Cost Estimates. *Journal of Cost Analysis and Parametrics, 6*(1), 43-74.

SAF/FMCE. (2012). Air Force Inflation Calculator.

Smirnoff, J., & Hicks, M. (2007). *The Impact of Economic Factors and Acquisition Reforms on the Cost of Defense Weapon Systems.* Elsevier Inc.

The White House. (2014). *Office of Management and Budget Historical Tables*. Retrieved January 13, 2014, from www.whitehouse.gov/OMB/budget/Historicals.

Tracy, S. P. (2005). Estimate at Completion: A Regression Approach to Earned Value. *MS Thesis, AFIT/GCA/ENC/05-04.* School of Engineering and Management, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2005: Print.

U.S. House of Representatives, 98th Congress (Version Date: 10/19/1984). *H.R. 5167, Department of Defense Authorization Act of 1985.* Washington: GPO

| REPORT DOCUMENTATION PAGE | *Form Approved*<br>*OMB No. 0704–0188* |
|---|---|

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704–0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202–4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

| | | |
|---|---|---|
| 1. REPORT DATE (DD–MM–YYYY)<br>27-03-2014 | 2. REPORT TYPE<br>Master's Thesis | 3. DATES COVERED *(From — To)*<br>Aug 2012 – Mar 2014 |

4. TITLE AND SUBTITLE
A Macro-Stochastic Approach to Improved Cost Estimation for Defense Acquisition Programs

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
DeNeve, Allen, J, Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology
Graduate School of
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT-ENV-14-M-20

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Dr. Alan Ashworth
Tri-Service Research Laboratory
4141 Petroleum Road
Fort Sam Houston, TX 78234
210-539-8504

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Distribution Statement A. Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

Inaccurate cost estimates are a recurrent problem for Department of Defense (DoD) acquisition programs, with cost overruns exceeding billions of dollars each year. These estimate errors hinder the ability of the DoD to assess the affordability of future programs and properly allocate resources to existing programs. In this research, the author employs a novel approach called "macro-stochastic" cost estimation for significantly reducing cost estimate errors in Major Defense Acquisition Programs (MDAPs). To achieve this reduction, the author first extracts and catalogs key programmatic data from 936 Selected Acquisition Reports. The author then analyzes historical trends in the data using mixed-model regression with high-level descriptive program parameters. Based on these trends, the model is found to reduce estimate errors by 18.7 percent on average, when applied to a randomly selected, historical cost estimate. However, the model is most beneficial when applied early in program life; when applied to the first cost estimate of each program in the database, the macro-stochastic technique reduces cost estimate error by over one-third. This statistically and economically significant reduction could potentially allow for reallocation of $6.25 billion, annually, if applied consistently to the DoD's portfolio of MDAPs.

15. SUBJECT TERMS
Cost estimation, macro-stochastic, cost growth, SAR, acquisition baseline

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON<br>LtCol Erin T. Ryan, AFIT/ENV |
|---|---|---|---|---|---|
| a. REPORT<br>U | b. ABSTRACT<br>U | c. THIS PAGE<br>U | UU | 109 | 19b. TELEPHONE NUMBER (Include Area Code)<br>(937) 785-3636 x3348 Erin.Ryan@afit.edu |

**Standard Form 298 (Rev. 8–98)**
***Prescribed by ANSI Std. Z39.18***